Biometric Storyboards:

A Games User Research Approach for Improving Qualitative Evaluations of Player Experience

Pejman Mirza-Babaei

Thesis submitted for Degree of Doctor of Philosophy

University of Sussex

August 2013

Declaration

I hereby declare that this thesis has not been and will not be, submitted in whole or in part to another University for the award of any other degree.

Signature:

University of Sussex

Thesis submitted by Pejman Mirza-Babaei for the Degree of Doctor of Philosophy

August 2013

Biometric Storyboards: A Games User Research Approach for Improving Qualitative Evaluations of Player Experience

Summary

Developing video games is an iterative and demanding process. It is difficult to achieve the goal of most video games — to be enjoyable, engaging and to create revenue for game developers — because of many hard-to-evaluate factors, such as the different ways players can interact with the game. Understanding how players behave during gameplay is of vital importance to developers and can be uncovered in user tests as part of game development. This can help developers to identify and resolve any potential problem areas before release, leading to a better player experience and possibly higher game review scores and sales. However, traditional user testing methods were developed for function and efficiency oriented applications. Hence, many traditional user testing methods cannot be applied in the same way for video game evaluation.

This thesis presents an investigation into the contributions of physiological measurements in user testing within games user research (GUR). GUR specifically studies the interaction between a game and users (players) with the aim to provide feedback for developers to help them to optimise the game design of their title. An evaluation technique called Biometric Storyboards is developed, which visualises the relationships between game events, player feedback and changes in a player's physiological state. Biometric Storyboards contributes to the field of human-computer interaction and GUR in three important areas: (1) visualising mixed-measures of player experience, (2) deconstructing game design by analysing game events and pace, (3) incremental improvement of classic user research techniques (such as interviews and physiological measurements).

These contributions are described in practical case studies, interviews with game developers and laboratory experiments. The results show this evaluation approach can enable games user researchers to increase the plausibility and persuasiveness of their reports and facilitate developers to better deliver their design goals. Biometric Storyboards is not aimed at replacing existing methods, but to extend them with mixed methods visualisations, to provide powerful tools for games user researchers and developers to better understand and communicate player needs, interactions and experiences. The contributions of this thesis are directly applicable for user researchers and game developers, as well as for researchers in user experience evaluation in entertainment systems.

Acknowledgements

I am very thankful and lucky to know and have worked with so many kind, smart, and generous people. Here are just a few who have helped me to make my dream possible:

First of all thanks to my supervisors for all their enthusiasm, time, ideas, and the opportunity they have given me. Dr. Graham McAllister: for teaching me about games user research, providing me with the opportunity to work with professional game developers and specially his vision on conducting applicable research has resonated with me. Dr. Lennart Nacke: for supporting me with physiological evaluation and experiment design, for hosting my visit at UOIT and always encouraging scientific rigor and using statistical analysis. Dr. Nick Collins: for being so supportive and patient, for motivating me over the final year of my Ph.D., and for helping me to develop an academic writing style. Prof. Geraldine Fitzpatrick: for teaching me HCI, her amazing lectures made me want to continue my studies in this domain, and many thanks for supporting me in the last years, not only with my Ph.D. research, but also helping me to develop my professional career in academia.

I would also like to show my appreciation to my Thesis Advisory Committee: Dr. Judith Good and Dr. Sam Hutton for their constructive feedback on my thesis, and thanks to my examiners: Dr. Anders Drachen and Dr. Paul Newbury for their valuable comments that helped me improve my work.

I have enjoyed my time at Sussex and have worked with some great people. I am grateful to all of the wonderful people at the Interactive System Group: Gareth R. White, Dr. Kate Howland, Jim Jakson, Ellie Martin, Chad Mckenny, Dr. Chris Keifer, Eric Harris, Dr. Lesley Axelrod, Edgar Cambranes, and Prof. Ben du Boulay. I would like to especially thank Ben for organising “surgery sessions” and providing me with the opportunity to discuss my research and get his constructive feedback, as well as his advice for starting my academic career.

I was lucky to have had the opportunity to spend six months as a visiting researcher at the University of Ontario Institute of Technology (UOIT), which would not have been possible without the understanding and support from Prof. Bernard Weiss (former Head of School of Engineering and Informatics), Prof. John Carroll (Head of Informatics Department) and Prof. Peter Cheng (Director of Doctoral Studies) to whom I am very thankful.

I am thankful to the members of the UOIT GAMER Lab for making me feel welcome; I would like to especially thank Dr. Bill Kapralos, Dr. Andrew Hogue and the Faculty of Business and IT administrators for facilitating my stay there.

I would like to thank Prof. Stephen Fairclough and Dr. Regan Mandryk for organising the Brain and Body Interface Workshop at CHI 2011. Thank you for some great discussions surrounding my work. I was also lucky to have the chance to present my work and get feedback from IGDA GUR SIG, CHI and DiGRA communities.

I was very fortunate to be able to work with some talented people, who contributed in the studies reported in the thesis: John Gregory for programming the Biometric Storyboards tool and also developing the *Matter of Second* game prototypes. Mark Knowels-Lee, Jason Avent, Jon Napier, Mina Tawadrous, Joel Lavigne, Brodie Stanfield, Christopher Zerebecki, Kei Turner and James Robb for providing me with their professional game design perspective. Emma Foley and Sebastian Long for their help in the analysis of observational data in the first study. Steve Bromley for his contribution to the second study. Mirza Beig for programming the physiological measurement software and also for making awesome videos and music for our CHI2013 paper presentation. Veronica Zammitto, Mirweis Sangin and Joerg Niesenhaus for co-organising the CHI 2013 GUR workshop with me. I would also like to thank Jodie Green for her great work on proofreading my thesis.

Thank you also to all my friends, especially Mohsen Fatorechi and Shahin Gheytanchi, with whom I shared the ups and downs of my Ph.D. journey, and Behzad Farzipour and his family for making me feel at home while I was in Canada.

I am very grateful and owe a lot to my family and Edina. Edina has provided me with so much support and encouragement through my Ph.D., and together with Maple, they have provided lots of fun and enough madness at home to keep me going whilst writing this thesis.

Publications

List of publications related to this thesis:

2013:

P1: **Mirza-Babaei, P.**, Nacke, L. E., Gregory, J., Collins, N., & Fitzpatrick, G. (2013). How does it play better?: exploring user testing and biometric storyboards in games user research. Presented at the CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. doi:10.1145/2470654.2466200. Paris, France.

P2: **Mirza-Babaei, P.**, Zammitto, V., Niesenhaus, J., Sangin, M., & Nacke, L. E. (2013). Games User Research: Practice, Methods, and Applications. In Proceedings the CHI EA '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. doi: 10.1145/2468356.2479651. Paris, France.

P3: McAllister, G., **Mirza-Babaei, P.,** & Avent, J. (2013). Improving gameplay with game-oriented and player-oriented metrics. In Game metrics: Maximizing the Value of Player Data (M. Seif El- Nasr, et al. Eds.) Springer. ISBN 978-1-4471-4768-8. DOI: 10.1007/978-1-4471-4769-5_26.

P4: Bromley, S., **Mirza-Babaei, P.**, McAllister, G., & Napier, J. (2013) Playing to Win? Measuring the Correlation Between Biometric Responses and Social Interaction in Co-located Social Gaming. In Multiplayer: The Social Aspect of Digital Gaming (T. Quandt and S. Kroeger, Eds.) Routledge. ISBN 978-0-415-82886-4.

2012:

P5: **Mirza-Babaei, P.**, Nacke, L. E., Fitzpatrick, G., White, G.R., McAllister, G., & Collins, N. (2012). Biometric Storyboards: Visualising Game User Research Data. In Proceedings the CHI EA '12: Extended Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. DOI: 10.1145/2223656.2223795, Austin, TX, USA .

P6: **Mirza-Babaei, P.**, Nacke, L.E. & McAllister, G. (2012). Biometric Storyboards: toward a better understanding of player experience. Presented at the CHI GUR Workshop 2012, Austin, TX, USA.

P7: **Mirza-Babaei, P.**, & Nacke, L.E. (2012). Biometric Storyboards: An Industry-Friendly Method for Evaluating Affect and User Experience in Games. Presented Research Note GRAND 2012, Montreal, Canada.

2011:

P8: **Mirza-Babaei, P.**, Long, S., Foley, E. & McAllister, G. (2011). Understanding the Contribution of Biometrics to Game User Research. Presented full research paper in Proceedings of the 5th DiGRA Conference 2011, Utrecht, The Netherland.

P9: **Mirza-Babaei, P.**, & McAllister, G. (2011). Biometric Storyboards: visualising meaningful gameplay events. Presented at the CHI 2011 Brain and Body Interface (BBI) Workshop. Vancouver, Canada.

P10: **Mirza-Babaei, P.**, & McAllister, G. (2011). Biometric Storyboards: visualising meaningful gameplay events. Presented at The 3rd Video Games Cultures and Future Interactive Entertainment Conference. Oxford, UK.

P11: **Mirza-Babaei, P.** (2011). Biometrics to Improve Methodologies on Understanding Player's Gameplay Experience. Presented at the 25th BCS Conference on Human-Computer Interaction British HCI (Doctoral Consortium). Newcastle Upon Tyne, UK.

2010:

P12: **Mirza-Babaei, P.**, & McAllister, G. (2010). Using physiological measures in conjunction with other UX approaches for better understanding of the player's gameplay experiences. Presented at the Games Research Methods Seminar 2010. University of Tampere. Finland.

Table of Contents

1 Introduction 1

1.1 Why Focus on Video Games? 1

1.2 What Problem This Thesis is Solving? 3

1.3 Working Definitions 5

1.4 Contributions 6

1.5 Outline of Thesis 7

1.6 Summary 10

2 Games User Research & Physiological Evaluation 12

2.1 Introduction 12

2.1.1 Video Games User Research 15

2.2 GUR Methods 17

2.2.1 Behavioural Observation 17

2.2.2 Think-aloud 18

2.2.3 Heuristic Evaluation 19

2.2.4 Questionnaires 19

2.2.5 Interviews 20

2.2.6 Metrics 21

2.3 Physiological Evaluation 23

2.3.1 Physiological Measurements 25

2.4 Mixed Methods 29

2.5 Summary 36

3 Mixed Methods Implementations 37

3.1 Triangulation of User Research Methods 37

3.1.1 Player's Self-assessment Diagrams 37

3.1.2 Player's Physiological Arousal to Structure Post-session Interview and Coding Gameplay Events 39

3.2 Study One: Using Physiological Arousal to Structure Post-session Interview 40

3.2.1 Data Collection and Setting 41

3.2.2 Results 44

3.2.3 A Closer Look 46

3.2.4 Discussion 48

3.3 Study Two: Physiological Arousal and Social Interaction Coding 52

3.3.1 Motivation *53*
3.3.2 Introduction *53*
3.3.3 Social Interaction *54*
3.3.4 Player Profiling *56*
3.3.5 The Study *56*
3.3.6 Results *62*
3.3.7 Discussion - What Does This Mean for Game Developers? *65*
3.4 Summary **69**

4 Biometric Storyboards: The Prototypes **70**
4.1 An Iterative Design Cycle **70**
4.2 Introduction **72**
4.2.1 Storytelling *72*
4.2.2 Datasets for BioSt Prototypes *75*
4.3 First Case Study **76**
4.3.1 Setting *76*
4.3.2 Overview of User Test Findings *77*
4.3.3 Biometric Storyboards First Prototype *77*
4.4 Second Case study **78**
4.4.1 Setting *78*
4.4.2 Overview of User Test Findings *79*
4.4.3 Biometric Storyboards Second Prototype *80*
4.5 Third Case Study **81**
4.5.1 Setting *81*
4.5.2 Overview of User Test Findings *81*
4.5.3 Biometric Storyboards Third Prototype *82*
4.6 Prototypes Evaluation **82**
4.6.1 Method *82*
4.6.2 Results *83*
4.6.3 Key Findings *85*
4.7 Discussion **87**
4.8 Conclusion **88**
4.9 Summary **89**

5 Biometric Storyboards: The Tool **90**
5.1 Introduction **90**

5.2 Interaction Steps **91**
5.3 Measuring Physiological Data **92**
5.3.1 Galvanic Skin Response *93*
5.3.2 Facial EMG *94*
5.3.3 Output Text File *95*
5.4 The BioSt Tool **95**
5.4.1 Designer's Intended Player Experience Graph *95*
5.4.2 Player's Input View *96*
5.4.3 GUR View *97*
5.5 Summary **100**

6 Evaluating Biometric Storyboards **101**
6.1 Introduction **101**
6.2 Related Work **102**
6.3 Overview of Evaluation **103**
6.4 The Game: Matter of Second (MoS) **103**
6.5 Phase1 - User Test Sessions **104**
6.6 Phase 2 – Developing Three Game Prototypes **105**
6.7 Phase 3 - Implementing Requested Changes **107**
6.8 Phase 4 - Experiment **107**
6.8.1 Experimental Procedure *107*
6.8.2 Participants *108*
6.9 Results **108**
6.9.1 Results from SUS, PANAS, and SAM Scales *108*
6.9.2 Results from Personal Preference Ratings *109*
6.9.3 Results from the Players' Interviews *110*
6.9.4 Results from the Game Designers' Interviews *112*
6.9.5 Results from the Game Programmer's Interview *112*
6.10 Discussion **113**
6.11 Conclusion **116**
6.12 Summary **116**

7 Discussion, Conclusion and Future work **117**
7.1 Summary **117**
7.1.1 Study One *117*
7.1.2 Study Two *117*

7.1.3 Case Studies: BioSt Prototypes Iteration *118*
7.1.4 Study Three: BioSt Prototypes Evaluation *118*
7.1.5 BioSt Tool *118*
7.1.6 Study Four: BioSt Evaluation *118*
7.2 Thesis Discussion **119**
7.3 Thesis Contributions **122**
7.3.1 Using Player's Physiological Measures to Structure Post Gameplay Interview *122*
7.3.2 Deconstructing Game Design by Analysing Pace and Events *123*
7.3.3 Using Player's Physiological Measures to Visualise Their Gameplay Experience *123*
7.3.4 BioSt Tool and Method for Analysing Player Experience *123*
7.3.5 Guidelines for Reporting GUR Findings *123*
7.3.6 Systematic Explanation on How GUR Help Improve Gameplay Experience *123*
7.4 Limitations and Future Work **124**
7.4.1 Including Game Analytics Data *124*
7.4.2 Improvement on BioSt Tool *125*
7.4.3 Further Study into Contributions of Biometrics in GUR *125*
7.4.4 Framework for Summative Evaluation of Player Experience *125*
7.4.5 Other Applications: *125*
7.5 Conclusion **125**

Bibliography **128**

Appendix 1: Abbreviations and Acronyms **142**
Appendix 2: Consent Form For S1, S2, S3 and Case Studies **144**
Appendix 3: Consent Form For S4 **145**
Appendix 4: Interview Schedule for S3 **148**
Appendix 5: Player's Demographics Questionnaire **149**
Appendix 6: Questionnaire used in S4 - Game's Features **151**
Appendix 7: Final Rating Questionnaire used in S4 **152**

List of Figures

Figure 1-1 Thesis contribution on better understanding of player experience 7

Figure 2-1 GUR studies, the focus of this thesis is on enhancing methods for applied GUR to provide feedback for a game under-development 16

Figure 2-2 Three components are involved in recording physiological metrics: external physical activity, internal emotional activity and internal cognitive activity (Nacke, 2013) 24

Figure 2-3 Russell, Weiss, & Mendelsohn, (1989) model of arousal and valence. The relationship between arousal/valance and physiological signals has since been applied to measure player experience. 26

Figure 3-1 Example of a player's self-assessment diagram for 30 minutes of gameplay 38

Figure 3-2 Example of a player's self-assessment diagram showing recalled events at the start and the end 38

Figure 3-3 Example of a peak in GSR signal 39

Figure 3-4 GUR studio – playroom at Sussex University 41

Figure 3-5 Example screenshot of the gameplay video 42

Figure 3-6 GSR sensors attached to ring and little fingers 43

Figure 3-7 Comparison of number of issues 46

Figure 3-8 The Buzz Controller 57

Figure 3-9 The social interaction recording tool interface, displayed recording a Shared History behaviour 61

Figure 3-10 Observation screen for S2 61

Figure 3-11 A Killer's self-assessment graph (player B2) 66

F1igure 4-1 Biometric Storyboards; top: first prototype, middle: second prototype, bottom: third prototype 70

Figure 4-2 BioSt design cycle 71

Figure 4-3 Example of story arc picture 73

Figure 4-4 Example of a story arc (left) compared to an example of a player's GSR over time (right) 73

Figure 4-5 Example of BioSt's first prototype 78

Figure 4-6 Key features of the BioSt second prototype 80

Figure 4-7 Biometric Storyboards second prototype. 80

Figure 4-8 Biometric Storyboards third prototype 82

Figure 4-9 Categorised interview results 83

Figure 4-10 Possible next iteration of BioSt 86

Figure 5-1 BioSt tool system design 92

Figure 5-2 GSR Sensors 93

Figure 5-3 Attachment of facial EMG sensors (Cacioppo et al., 2007). 94

Figure 5-4 Screenshot of Designer's graph draw mode 96

Figure 5-5 Player's Input screen 96

Figure 5-6 GUR screen view 97

Figure 5-7 Percentage of smiling (green) and frown (red) muscles activity in one level 98

Figure 5-8 Intended player experience graph (representing what designers think exciting gameplay moments are) showing game segments, times and key nodes 99

Figure 5-9 Single player's data graph in GUR view synced based on the frame counter timestamp 99

Figure 5-10 GUR aggregated player experience graph, indicating areas of difficulty and average time spent in each segment 99

Figure 6-1 The overview of evaluation 103

Figure 6-2 Screenshot of MoS level 1 104

Figure 6-3 Significant mean (CI: 95%) PA rating from PANAS 109

Figure 6-4 Significant average (CI: 95%) SAM Pleasure Rating 109

Figure 6-5 Significant average (CI: 95%) preference ratings 110

Figure 7-1 Prototype idea to include map of player avatar's death locations (red dots) into BioSt 124

Figure 7-2 Current methods of GUR 126

Figure 7-3 BioSt combines objective/subjective and qualitative/quantitative evaluation methods in a single mixed method 126

List of Tables

Table 1-1 Overview of the thesis chapters with regard to BioSt development process, related studies and resulting publications 8

Table 2-1 Game versus productivity applications (Pagulayan et al., 2003) 14

Table 2-2 Criteria for applied GUR studies (Fulton et al., 2012) 17

Table 3-1 Participants' information: (1) ID, (2) Age, (3) Favourite games type, (4) Preferred platform, (5) Frequent gaming, (6) preferred play condition. 42

Table 3-2 Issue categories obtained from Desurvire & Wiberg (2009) 45

Table 3-3 Showing Voida & Greenberg (2009) categories of social interaction behaviour, and the adapted categories (Bromley, 2011) used in this study. 55

Table 3-4 Session data showing the players and rounds present in each game 60

Table 3-5 Table showing the percentage of the session's total interactions in which each type of social interaction was noted. 63

Table 6-1 Game conditions 107

Table 7-1 Evaluating BioSt based on criteria for applied GUR studies 121

1 Introduction

This thesis presents an investigation into the contribution of physiological measurements in Games User Research (GUR) within the larger context of HCI (Human-Computer Interaction). GUR professionals specifically study the interaction between a game and its players (often referred to as users in HCI), to provide feedback for game developers to help them to optimise the experience their game provides. A better understanding of players and their interactions within a game (i.e., gameplay) will enable user researchers to increase the plausibility and persuasiveness of their reports.

In this thesis, approaches are introduced which use physiological measurements to better understand player experience. These approaches are not aimed at replacing existing user research methods, but to extend and establish a novel combination of existing methods, which could provide powerful tools for games user researchers (GURs) to better understand players and their experience within a game. Previous researchers, who have been applying physiological measurement in game evaluation, focused on providing summative evaluation, or to utilise these measures as independent variables for comparison of different experimental conditions. However, in this thesis the focus is on the methodological development of physiological measures in providing formative evaluation during game development cycles. This is a novel contribution in GUR and HCI, which is described by way of case studies in professional practice, interviews with game developers and laboratory experiments. This thesis tackles an important problem in an applied GUR setting, which is to improve *physiological-based approaches for evaluation of player experience* by developing a quick, cheap and easy-to-understand method that integrates these data into GUR. The method is referred to as Biometric Storyboards (BioSt) and developed through an iterative design and evaluation process (see Table 1-1) and covers the areas of:

- Visualising components of player experience
- Identifying key gameplay moments
- Deconstructing gameplay design

The upcoming sections review the research context; they further elucidate the research questions and contributions of this thesis, before laying out the content summaries of the thesis chapters.

1.1 Why Focus on Video Games?

Designing and developing video games is often a challenging and demanding process. The overall aim of developing a game, which is enjoyable and rewarding to play for everyone as well as making profit for the game developers, is a complex one due to the diversity of players

who may potentially interact with the game. Understanding how players interact and behave during gameplay is of vital importance to developers. An accurate understanding of gameplay experience during development of a game can help to identify and resolve any potential problem areas before release, leading to a better player experience and arguably greater game review scores and sales (Fullerton, 2008).

Over the past decades the video games market has rapidly increased as part of the entertainment industry. It is predicted that video games will be the fastest-growing form of media over the next few years, with sales rising to $82 billion by 2015 (The Economist, 2011a).

Beside the rapidly growing market, there are other factors that make video games an increasingly popular topic for research. I am describing examples of these factors briefly below to explain why the understanding of users and their interactions with games are critical to designing a successful game:

Demographic: Video games have attracted new audiences, such as women and older adults, who have had no prior interest in games. According to the Entertainment Software Association the average age of players in America is 37, and 42% of them are female (The Economist, 2011a).

Platforms: The video games industry is experiencing changes in how players interact with games. For example, Nintendo's Wii provides simple design, intuitive motion-sensitive controllers and non-complex games (based on sports and puzzles) designed for accessibility to non-gamers (The Economist, 2011b). On the other side, the increasing computing power of tablets and mobile phones provides new gaming platforms and interaction methods (e.g., touch) that attract games (and gamers) that have less time and dedication than a console title.

The Internet: The Internet also has a crucial part in today's video games industry both as a gaming, advertising and a marketing platform. The Internet as a games platform has its own specification, providing a new sociable experience for gamers located in different parts of the world. Social networking websites, such as Facebook and Google+ are popular portals for video gaming and provide their own development tools to specifically integrate their social features into the online games. For example, video games make up half of the 40 top grossing applications on Facebook, and social-networking game development company *Zynga* earns revenues of around $850m a year (The Economist, 2011b).

Distribution*:* The Internet enables game developers to sell their product without the need of traditional publishers. Online services such as Steam[1] make it possible for small, independent

[1] www.steam.com

developers to join the market. Similarly, the marketing model in Apple's app store creates a locked-down, competitive market so that only highly rated (or highly downloaded) games are promoted to users and survive the market. On the other side, more and more games are free to play (F2P) or cheap, but the virtual goods available in these games cost real money. For example, in "*FarmVille*" (Zynga's popular Facebook farming simulator game) players can earn coins to spend on crops, livestock or farm equipment by playing the game, or they can buy them with real cash. The microtransactions generate the bulk of Zynga's revenues (The Economist, 2011c). Many other similar games are being created following the F2P model that often aim at enjoyment, frustration and friction towards monetisation. Therefore, the development profit directly depends on the game being consistently engaging to play.

As mentioned above, there have been many changes in video games development in recent years, including new business models, widening player demographics and new controller interfaces. These present opportunities, but also additional uncertainties and —in combination with escalating design and development costs often for large-scale titles— developers are focused on ensuring every game is a success. The opportunity is the wider market. The challenge is that new demographics or platforms require a deeper understanding of players. Generally, the industry is moving from one player stereotype (e.g., the stereotypical image of a teenage boy, who plays many hours per day on a game console) to ubiquitous gamers across multiple homes and mobile platforms. The key point is that modern video games offer different ways of interaction to deliver a better player experience. Therefore, the demand for research studies dealing with users and their interaction with video games has grown in recent years.

1.2 What Problem This Thesis is Solving?

One approach that has increasingly been applied to improve gameplay experience and make sure it meets designers' criteria is GUR. GUR is an emerging field that has borrowed methods, such as behavioural observation, interview, questionnaire and heuristic evaluation, from HCI and psychology. Although HCI methods have made progress in understanding the usability of productivity applications, applying the current methods to identify user experience (UX) issues in video games is still a challenge for researchers. Due to the specific characteristics of video games (such as emotion and intentional challenge), most of the well-established user research methods (especially those based on self-reporting) cannot be used the same way for formative evaluation of player experience. GUR methods have been adapted and evolved with the aim to provide a mixture of qualitative and quantitative approaches for evaluators to choose from depending on their research question. However, identifying the effective mixture of these methods and mapping (or blending) the results from each into a meaningful, actionable, adequate, convincing and easy-to-understand user test (UT) report to be presented to game

developers is one of the current challenges facing GUR. As discussed by Fulton, Ambinder, & Hopson (2012) the challenges of applied GUR are: 1) GUR methods should provide researchers with *accurate* results that reflect on user testing's assumptions. 2) GUR methods should provide researchers with *specific and precise* results that also need to be 3) *actionable and applicable*. 4) When conducting the studies, analysing the data and reporting the findings, all need to be completed in a *time frame* that suits the game development cycle. 5) Value added to a game by user test findings must *return the cost* of running the user test. 6) Presented results need to motivate game designers to take action on them.

Another challenge facing GUR is to better understand players' gameplay experience, but the usual methods of determining the players' experience are based on self-reporting methods. Although these are relatively easy to conduct (e.g., questionnaires) and can potentially provide a rich source of data (e.g., interviews), they are relying on player's ability to recall and explain their experience.

In order to address these issues, GURs are using a number of innovative techniques to enhance GUR studies to better understand players' experience, and new methods of communicating user test findings to the game development team. As such, there is an increased interest in research concerning use of physiological measurements as an addition to traditional methods for evaluating user experience in video games.

Physiological measurements (also called biometrics) refer to the capture and analysis of signals directly from the user's body using skin-contact sensors that show how different users react to events or stimuli on the screen. Since physiological evaluation measures often visceral biological responses, they are instinctive and hard to fake. Researchers have used physiological measurements to evaluate an emotional experience, however, in the game industry, apart from few cases within big publishers (i.e. Zammitto, 2011; Ambinder, 2011). Hence, these measures have not been widely used to provide formative feedback during game development. The complexities of physiological-based approaches (i.e. knowledge needed and accessibility of applicable methods) often outweigh the non-established benefits of these approaches over traditional user research methods, such as observation and interviews. Thus physiological-based evaluations are less utilised among practitioners in medium or small-scale studios.

Based on previous research on the use of physiological techniques (e.g. Mandryk, 2008; Nacke, 2009; Hazlett, 2008), this thesis explores whether capturing, measuring and analysing player's physiological responses can provide GURs with new mixed methods. This extra data set can be used by GURs in triangulation with other evaluation methods (e.g., subject reports and video analysis) to enhance understanding of players and their experience with a video game.

This thesis aims to make a contribution of methodological improvement and extension of mixed methods in GUR, combining physiological measures with other user research methods in a reporting tool named Biometric Storyboards. The goal is to develop an approach that supports GURs to effectively identify and communicate playability issues with game development team.

The motivation of this research is that optimising user experience methodologies to better understand the target users are of benefit to video game development, with **the following research questions to be answered**:

1. What are the limitations of classic user research methods in GUR?
2. What are the limitations and potential contributions of physiological measures in GUR?
3. How can physiological evaluation data be presented in a comprehensible format for providing formative feedback in video games development?

One of the motivations of this thesis was the Sussex University's Human-Computer Technology group's[2] interest in developing user-centred approaches to better understand users by applying human-centred technologies. In particular, for this research, this means using GUR techniques to improve understanding of users to provide feedback for game developers. Previous research from the group (McAllister & White, 2010) noted that the traditional game development process and user testing methodology are suffering from less effective methods of evaluation. Improving on these limitations is a main aim of the research presented in this thesis.

Biometric Storyboards consists of foundational methodologies (objective metrics and subjective evaluation to drive qualitative results) which were iterated and developed to fit into the overall goal of this Ph.D. research, these methodologies are:

1. Physiological recording
2. Self-report
3. Observation
4. Event-coding

1.3 Working Definitions

Video games are an interdisciplinary research area, providing new research challenges to many fields. For example, research can range from a humanistic, ludological perspective – which

[2] The Sussex University's HCT Group overall researchers objectives are to develop frameworks for understanding how people interact with and communicate through technology; also to apply this understanding to develop and support innovation. The group's interest is in developing methodologies and frameworks that incorporate direct research with users into different settings. These have been applied to ranges of technologies and domains including: video games (e.g. Vertical Slice (Hakner, 2009)), learning (e.g. (Frauenberger, Good, Alcorn, & Pain, 2012)), creativity (e.g. (Good, Howland, & Nicholson, 2010)), supporting special users (e.g.(Balaam, Fitzpatrick, Good, & Harris, 2011)), ageing in place and healthcare (e.g.(Axelrod et al., 2011)).

focuses on establishing design vocabularies (e.g. (Costikyan, 2002)) – to a technical perspective based in computer sciences – which focuses on creating new game engines to improve technical quality (e.g., rendering, AI, performance, physics) of games. Of course there is much literature in these areas, which cannot be covered in this thesis. While the focus is on games user research methods, the scope of this thesis is to study the *interaction* between a *game* and its *users* (and between users), with the aim of providing *formative feedback* for game developers to help them to optimise the *experience* of their title in development.

Game user experience (UX) or player experience (PX) research*;* while many researchers would see game UX research as the same as GUR, for the context of this thesis I see a difference. A game UX researcher would also study the *interaction* between a *game* and its *users*, but their aim is to prove (validate) theories or provide general design guidelines. In this thesis I sometimes use the *player experience* (PX) term, this is only to avoid repeating UX several times (i.e., in this thesis PX=UX). Although I think the PX term is more common in ludological studies.

User test (UT), play test or usability, playability*;* for this research our participants are simply *users* that interact with a game (not *players* in ludological terms); again, this is only to keep the focus of the work on the HCI and Computer Science perspective. Terms like "user test" or "play test" have a similar meaning for me in this thesis, although researchers in other field may see a difference in them.

1.4 Contributions

The ultimate aim of this thesis is to enhance the understanding surrounding player experience as part of the formative design process in video games development cycle. To answer this, this thesis focuses on an investigation into the contributions of physiological measurements in GUR, with the goal of developing an applicable method based on the requirements of game developers and GURs.

This thesis (and its resulting publications) contributes to the field of HCI and GUR by proposing the Biometric Storyboards evaluation method, which uses physiological measurements with storyboarding techniques to visualise a meaningful relationship between game events and changes in player's physiological state. As stated earlier, BioSt itself is a novel combination of existing data gathering techniques all developed as part of this research. The results of this thesis show these evaluation methods would enable GURs to have a better understanding of affecting player's experiences, increase their confidence in reporting and convincing game development teams of the existence of playability issues. This deep understanding of players would ultimately enable developers to better deliver their intended experience design.

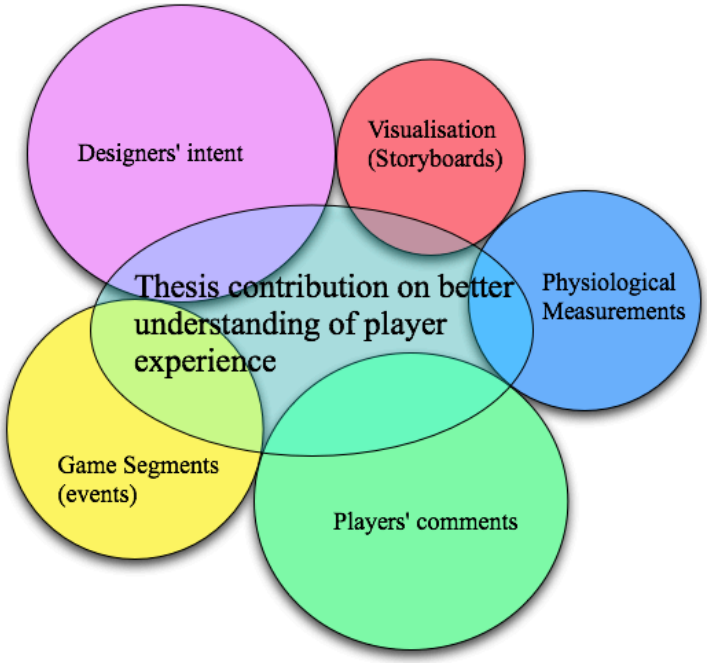


Figure 1-1 Thesis contribution on better understanding of player experience

The methods discussed in this thesis make the following contributions (see Figure 1-1):

1. Visualising player experience in a game.
2. Deconstructing game design by analysing events and pace.
3. Incremental improvement of classic user research techniques (such as self-report, storyboards, physiological analysis).

1.5 Outline of Thesis

As shown in Table 1-1 this thesis contains seven studies (three of them are explanatory case studies). These studies are presented in Chapters 3, 4 and 6 where each chapter provides: a relevant literature review (this is in addition to the thesis literature review in Chapter 2 and particularly focuses on research questions related to each study); study settings; results; discussion; limitations and conclusion. Chapter 7 provides a discussion, which reflects the thesis as a whole, posits the thesis research questions and contributions, as well as outlining questions for future work. A detailed outline of the thesis chapters are as follow:

Chapter 2: Games User Research & Physiological Evaluation

Chapter 2 has two main sections; the first section provides an overview on video games user research that includes discussion on why good gameplay experience is not always achieved. The main body of this section talks about existing user research methods (such as observation, interview, think-aloud, questionnaire and heuristic evaluation) and their advantages and limitations. It provides discussion on how these methods have been adapted for GUR, how some have been successfully applied, and where current methods have limitations in the

understanding of user experience in video games. This section concludes with why optimisation of these methods is needed to better suit GURs requirements, and why physiological measures can be used in this combination to increase understanding of user experience in video games development.

The second section of Chapter 2 provides an overview of the physiological concepts used in this research. This includes an explanation of physiological measures (mainly on GSR and facial EMG), issues and limitations associated with measuring physiological responses. This is followed by an example of studies using physiological measures for user experience evaluation in video games research.

The rest of the thesis describes my journey through creating and evaluating Biometric Storyboards. I describe three mixed methods; each provides a dataset used to create BioSt. I explain how and why each of these data sets contributes to the generation of BioSt. This includes: a description of two studies **(S1, S2)** in Chapter 3; three case studies with game studios on their under development title and one interview Study **(S3)** with six game developers in Chapter 4; and one final study in laboratory setting **(S4)** in Chapter 6. The knowledge gained from these studies helped me to create, iterate, develop and investigate the usefulness of Biometric Storyboards and its foundational mixed methods as result of this Ph.D. research. The result of these studies was published and presented in various academic conferences (See the list publications).

Chapter	**Title/Study**	**Main publication(s)**
Chapter 1	Introduction	-
Chapter 2	Games User Research & Physiological Evaluation	-
Chapter 3	Mixed Methods Implementations Study 1 | Study 2	*P12, P11, P8, P4*
Chapter 4	Biometric Storyboards: The Prototypes Case study 1 | Case study 2 | Case study 3 Interview Study 3	*P11, P9, P5, P3*
Chapter 5	Biometric Storyboards: The Tool	-
Chapter 6	Evaluating Biometric Storyboards Study 4	*P1*
Chapter 7	Discussion, Conclusion, Future Work	-

Table 1-1 Overview of the thesis chapters with regard to BioSt development process, related studies and resulting publications

Chapter 3: Mixed Methods Implementations

In Chapter 3 explains BioSt's three data sets and their foundational mixed methods. Each of these mixed methods is created as a novel approach resulting from triangulation of existing user research methods and explained by two studies with the aim of presenting detailed examples on how they can be applied. Section 3.1.1 describes players' self-assessment diagrams; section 3.1.2 details two mixed methods that developed from utilising players' physiological measurements in conjunction with post-session interview and observational behaviour coding for better understanding of player in game behaviour.

By applying these mixed methods in two studies, this chapter also discusses how the results from each of these mixed methods advances HCI techniques for games user research, as well as their contributions as a data set to implement BioSt. Section 3.2 details Study **S1** and demonstrates two sets of results, firstly a quantitative set as a comparison of the proposed mixed methods with traditional observation-based evaluation (Mirza-Babaei, Long, Foley, & McAllister, 2011) and second a qualitative set to explore how this approach can help to better understand the user's experience. This study aims to explore a contribution of using physiological measures for identifying usability and user experience issues in GUR studies. Then, section 3.3 describes the Study **S2** where physiological measurements and observational and verbal social interaction coding were used for better understanding of player gameplay behaviour. The effectiveness of this approach has been discussed further in the game design cycle (Bromley, Mirza-Babaei, McAllister, & Napier, 2013). The knowledge gained from these two studies has been used to further increase our understanding of the contribution of physiological measures in GUR.

Chapter 4: Biometric Storyboards: The prototypes

Chapter 4 introduces the Biometric Storyboards technique, where player's physiological measurement and storyboarding are used to develop an evaluation tool to better understand and communicate the player's experience. The development of this technique went over a number of iterations based on three case studies and one interview Study **(S3)**. The result of each Case study helped to evaluate the design of each prototype of BioSt and iterate it. The first and second prototypes were as a result of the first two Case studies (Mirza-Babaei & McAllister, 2011a). A further development of the technique is explained in third Case study that led to a third prototype (Mirza-Babaei & McAllister, 2011b). These three prototypes have been evaluated in the interview Study **(S3)** with six prominent game developers (Mirza-Babaei et al., 2012).

Chapter 5: Biometric Storyboards: The tool

The result from the prototypes' evaluation provides a base to develop an application that facilitates generating BioSt. This chapter discusses the BioSt tool's feature list (based on the evaluation results) and provides in-depth development details on the BioSt visualisation tool.

Chapter 6: Evaluating Biometric Storyboards

This chapter presents a study demonstrating how a classic UT and a BioSt UT both help designers create a better gameplay experience. The study's setting, conditions and results are explained in this chapter. In addition, this chapter shows that BioSt can produce high gameplay quality and visuals in comparison to designing without UTs, and that classic UTs do not provide this significant advantage. The results reported in this chapter support the idea that BioSt provides more focused and nuanced game design improvement (Mirza-Babaei, Nacke, Gregory, Collins, & Fitzpatrick, 2013).

Chapter 7: Discussion, Conclusion and Future Work

Chapter 7 provides a summary discussion and conclusion of the thesis and revised research questions considering the findings across the thesis chapters, as well as areas for future works in this domain. It should be noted that each study presented in each chapter is followed by its relevant discussion and conclusion sections, where they reflect on the process of the study and challenges of formative rather than summative evaluation, and argues for the contributions and repositions each study's research objectives.

These chapters are followed by references and 7 appendixes. Appendix 1 lists the abbreviations used in this thesis. Appendix 2 includes a copy of the consent form for studies S1, S2, S3 and the Case studies where ethics approval was gained and conducted at the University of Sussex. Appendix 3 includes a copy of the consent form for Study S4 where ethics approval was gained and conducted at the University of Ontario Institute of Technology (UOIT). Appendix 4 provides the interview schedule for S3. Appendix 5 provides a copy of player's demographics questionnaire used in the thesis. Appendix 6 provides a copy of the questionnaire on the game's features used in S4. Appendix 7 provides a copy of the questionnaire used for comparing game prototypes used in S4.

1.6 Summary

Video games are one of the most popular entertainment activities. Game developers and publishers are using emerging technology to deliver better video games. One aspect of improving video games is to better understand the target users (players). Methods from HCI allow GURs to better understand the users and their experience with video games. This information would help game designers to optimise the design of video games. One of the

methodological challenges for games user researchers is to be able to collect data from players without interrupting their gameplay. The need for continuous and unconscious methods for collecting data from users has grown in the last few years. Utilising physiological evaluation techniques is therefore becoming a more popular method together with interviews for games user research. This thesis contributes to the growing body of user research knowledge by detailing mixed methods approaches on utilising a player's psychological measurements in order to better understand users' experiences in video games.

The next chapter provides the background of the research, including reviewing relevant existing literatures on games user research methodologies, physiological sensors and measurements, particularly in HCI and user experience research.

2 Games User Research & Physiological Evaluation

This chapter explores the background literature to the main contribution of this thesis – Biometric Storyboards as a GUR method. This chapter starts by exploring challenges in game development in delivering a good gameplay experience. It discusses research in video games usability and user experience. Then it looks at classic user research methods that are often used in GUR, and importantly, how user researchers can apply a triangulation of these methods (mixed methods) to better understand players' behaviour and their experience.

This chapter also discusses issues related to gameplay analysis using classic approaches, and argues for adaptation of physiological measures for games user research based on previous studies that applied these measures in conjunction with classic user research resources. Furthermore, this chapter also looks at physiological measurements and evaluations; to explore the limitations and contributions of these measures in better understanding player behaviour, and goes into details of the two physiological measurements that are used in this thesis (GSR and facial EMG).

2.1 Introduction

Video games are art forms as much as they are new interactive media, often pioneering novel user interfaces and exploring new target groups. Thus, it is complex to evaluate how design decisions affect player experience. Evaluating game design frequently relies on the informal skills which game designers acquire on the job. Game design itself often follows the prescriptions and rules developed by people with considerable experience on the job (e.g., Formal Abstract Design Tools (Church, 1999)). Academic research also often takes a systematic stance to understanding game design, focused on observing, measuring and testing player reactions (e.g., prescribing design heuristics (Sweetser & Wyeth, 2005)). GUR lies somewhere between these approaches, aiming to improve player experience by providing sufficient information about gameplay for designers to draw the best conclusion possible for improving the experience of their games. While traditional video game testing has focused on the improvement of software (e.g., bug tracking and quality assurance), as part of GUR it has become more common to run user testing to improve a game's design (Isbister & Schaffer, 2008). Hence, user testing (UT) is becoming an essential part of most game's iterative development cycle.

Many researchers have studied game design concepts and gameplay criteria. For example, to investigate what components make a gameplay more fun and enjoyable (Malone, 1981; Sweetser & Wyeth, 2005). However, using data based on players' interaction with the game for evaluating and explaining their gameplay experience is relatively new.

Although classic user evaluation methods (such as observation or self report techniques) have made progress towards understanding the usability of productivity applications and websites, the specific characteristics of video games (see Table 2-1) mean that many established methods of user research cannot be applied in the same way for video games evaluation. For example, frustration is allowed in games as long as it is part of the game design and embedded in the game loop. Therefore, optimisation of evaluation methods suitable for GUR is an important topic for the HCI and GUR community. This optimisation often consists of a novel combination (i.e., mixed methods) of existing user research methods (such as RITE or TRUE which are discussed further in section 2.4) adapted suitable for requirements of GUR studies (see Table 2-2).

Research on a wider range of applications of technology and non-tangible factors such as feelings or experience has become an integrated part of expanding HCI research (the third wave of HCI) (Bødker, 2006). This offers a promising perspective for video games research as both fields share the mutual objective of evaluating affect and user experience between interactive technology and the humans using it.

Early works in the domain of video games and HCI evaluation emphasised a framework for designing instructional environments (e.g. (Malone, 1984)) or how user-centred design (UCD) methods could be applied into the video games industry, and the differences between games and productivity applications. For example Pagulayan, Keeker, Wixon, & Romero (2003), discussed ten areas that highlight these difference (see Table 2-1). Hence most classic UT approaches (such as questionnaires, interviews and focus groups, as well as observational video analysis reports) had to be refined and structured according to these differences.

Games vs. productivity applications	**Examples**
Process vs. results	The purpose of gaming is usually in the process of playing not in the final result.
Defining goals vs. importing goals	Games (or gamers) usually define their own goals, or how to reach a game's goal. However, in productivity applications the goals are usually defined by external factors.
Few alternative vs. many alternatives	Games are encouraged to support alternative choices to reach the overall goal, whereas choices are usually limited in productivity applications.
Being consistent vs. generating variety	Games are designed to provide a variety of experience, however productive applications are meant to be consistent in the user experience.
Imposing constraints vs. removing or structuring constrains	Game designers intentionally embed constraints into the game loop, but productivity applications aim to minimise constrains.
Function vs. Mood	Productivity applications are built around functionality, but games set out to create mood (e.g., the use of sound or music)
View of outcome vs. view of world	Gamers usually play a role in a game world. For example: in first-person shooter (FPS) games, but productivity applications rarely have a point of view.
Organisation as buyer vs. individual as buyer	Individuals usually buy games, but productivity applications are often bought by organisations.
Form follows function vs. function follows form	Gamers tend to welcome innovation but users of productivity applications tend to be cautious about adopting innovation.
Standard input devices vs. novel input devices	Games usually explore possibilities to use novel input methods (such as motion captures or biofeedback) in addition to standard input devices, on the other side productivity application mostly use a mouse and keyboard.

Table 2-1 Game versus productivity applications (Pagulayan et al., 2003)

Different approaches have been proposed to study user experience and the relationship between products and users. For example, Hassenzahl (2005) explored how users' actual experiences can differ from intended experiences. Hunicke, LeBlanc, & Zubek (2004) looked at how both designers' and players' perspectives need to be considered for game design and study. Nacke (2010) reviewed how the dynamic context of the player can change the game experience over time. Nacke & Drachen (2011) discussed three abstractions in the gameplay experience model:

the technical game *system*, the perceptive and operational player *actions*, and the *context* of the player in a moment of time. The aim is to make the impacts of the gameplay experience understandable and possibly quantifiable. This thesis evaluates players' experiences in comparison to game designers' intended experience, for example a negative experience may be intentional and allowed if it is part of the game design loop. This view reflects on the development of BioSt where the evaluation result (in Chapter 4) shows benefits of the tool if player experience is compared to designers' intended experience.

Studying user experience in games is complex, as many aspects of gameplay would influence it, including game mechanics, the context of the gameplay, previous experiences, preferences and perspective of the player. While many approaches are being developed and used as part of a mixed methods to understand game experience, it is still a challenge to capture and communicate player experience in a way that complements phasic quantitative data collection, such as game metrics or physiological measurements.

2.1.1 Video Games User Research

During the early years of GUR, the adaptations of evaluation methods (generally usability and productivity evaluation) to video games have provided an undeniable advancement to the field (e.g. (Pagulayan et al., 2003; Lazzaro & Keeker, 2004)). However, the variety of adaptations and the lack of consistency in variable-manipulation could be seen as shortfalls in assessing methodological efficiency and reliability. This situation reverberates in industry settings, where methodological optimisation has a direct impact on the inclusion of GUR in the development process and in return on investment (ROI) (Mirza-Babaei, Zammitto, Niesenhaus, Sangin, & Nacke, 2013). Thus, developing methodologies suitable for GUR and the game development cycle is in high interest of GURs.

Generally GUR studies can be conducted with two overall aims: one is to generate design guidelines and best practices, where it usually consists of a summative evaluation on often released games. The aim of this evaluation is to identify and explore what makes games fun, engaging or lead to a positive experience. This knowledge can be used to produce design guidelines with a focus on specific genre. Alternatively, GURs undertake research to perform a formative evaluation often on a game currently under development; the main aim of this type of study is to provide information for the game developers on how players perceived their intended design. The findings of studies of this type are usually reported back to the game developers; highlighting gameplay issues based on players behaviour, emotion, opinion and feedback. The focus of this thesis is on methodologies for conducting formative evaluation in GUR (I often refer to this as an *applied* GUR, see Figure 2-1).

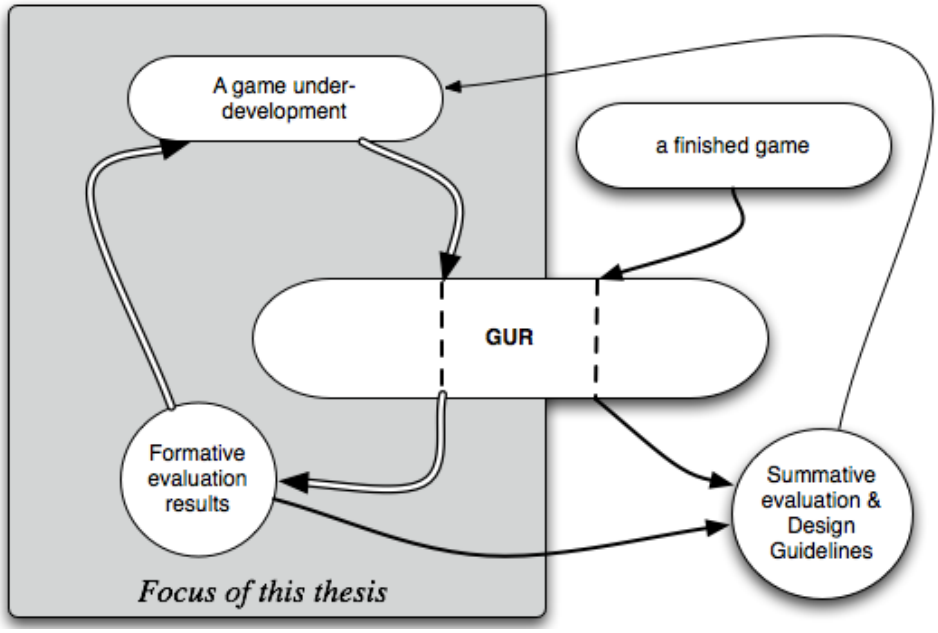


Figure 2-1 GUR studies, the focus of this thesis is on enhancing methods for applied GUR to provide feedback for a game under-development

The GUR report usually contains high-level material such as information on player behaviour and experiences, as well as more specific issues such as to identify game blockers, confusing language or concepts. Both aim to help game developers iterate their design to provide a better experience for players, which is played in the manner of their design intent.

There are many similarities between formative and summative evaluations in GUR, both areas are overlapping in many cases, however the goal of summative evaluation is usually to answer hypothesis and generate general guidelines and best practices that can apply to a wide variety of games. However, in formative evaluation the emphasis is on including players' feedback to improve the particular game's design and implementation. Here the main goal is to see if players interact or experience the game in the same way as the designers intended. As well as to remove any potential blockers and provide a smooth gameplay experience.

At the GUR summit 2012, Fulton, Ambinder, & Hopson (2012) discussed an evaluation framework to conduct GUR studies with a focus on formative evaluation. Table 2-2 summarises these criteria, which shows evaluation methodologies (the focus of this thesis) have a significant influences for the success of every GUR study. These criteria are the *heart* of the methods created and developed as part of this thesis.

Representative	Selected methods and recruited participants must correctly reflect on a UT's needs and outcomes.
Accurate	Results (analysed based on selected methods) should reflect on a UT's assumptions and supported by data (multiple source of data).
Specific	UT and methods selected for conducting the test needs to deliver precise and specific results (e.g. saying a game is not good without indicating why and where the problems are).
Timely	UT findings should be delivered in a time frame that matches the game development cycle. This reflects on selected methods, amount of data gathered and analysis approaches.
Cost-effective	Value added to a game by UT findings must return a cost of running the user test (which depends on selected methods and analysis approaches).
Actionable	UT needs to deliver actionable and applicable results. The quality of results directly effected by chosen methods and analysis approaches.
Motivational	Presented results (analysed based on selected methods) should motivate game designers to take action on them. Game designers should believe in and fully understand the results.

Table 2-2 Criteria for applied GUR studies (Fulton et al., 2012)

2.2 GUR Methods

This section looks into GUR methods and resources to better understand advantages and limitations of each approach, as well as to explain how studies conducted in this thesis utilised them.

2.2.1 Behavioural Observation

Behavioural observation logs are a primary resource in video GUR. Observation can provide a basis for a detailed analysis of usability (Blythe, Overbeeke, Monk, & Wright, 2004), and fun and game experience (Poels, de Kort, & Ijsselsteijn, 2007). Observation involves watching the player interact with the game and picking up cues from their gameplay behaviour, facial expressions, body language and social interaction in collocated gaming. A major benefit to using observation sessions is that they are relatively easy to conduct and can potentially provide a rich source of data; for these reasons observation is sometimes referred to as a core GUR method.

Observing potential players interacting with a game can be very useful for the game designers to see how players interact with the game; for example, to find out about blockers and unintended interactions. However, whilst analysis can be performed *'live'*, understanding behaviour requires precise interpretation and, unless the video data is captured and reviewed, important

events may be missed by researchers. Studying observational data as an indication of human experience is a lengthy and difficult process, which must be undertaken with great care to avoid biasing the result (Marshall & Rossman, 2010).

Observation is often applied in conjunction with other user research methods, as using only observation data would not always reliably answer *why* players performed specific behaviour or how they felt by doing that. Similar to other user research methods, observation can also be intrusive as players may feel pressure by knowing they are being watched. Therefore it is ideal if observation takes place in a room setting that simulates a natural playing environment. Players should also receive briefing about the purpose of the session. For example, that they should play the game as they naturally do and that it is about testing the game not testing how good they are at playing games. It is also recommended that players should play in separate observation rooms and not be distracted or interrupted in the process. In order to reduce bias on subjective interpretation of observational data, usually more than one observer conducts the session.

Observation is one of the core methods utilised as a data source for creating BioSt and other studies presented in this thesis. S1 looks at the type of potential issues that can be uncovered by using an observational based approach, which was compared to a physiological based approach. To enhance understanding of player experience, S2 proposes an approach from a triangulation of players' coded social interactions (from observing their gameplay and verbal comments) with physiological measurements.

2.2.2 Think-aloud

A commonly used extension to observation is to think-aloud or verbal reporting, which involves users describing their actions, feelings and motivations during the test session. Think-aloud was first develop as a technique for products' usability testing (Lewis & Mack, 1982) and aims to get inside the users' thinking processes *'in the moment'*, potentially revealing unobservable details and providing researchers with immediate feedback.

Although think-aloud is also a common user research approach, it is arguable that it cannot effectively be used within user testing sessions because of the disturbance to the player and ultimately the impact it has on gameplay (Nielsen, 1992). Think-aloud could, however, be seen as an add-on approach when conducting behavioural observation, as it would add players' thoughts to the observation data. It needs to be unprompted and participants need to say their thoughts aloud whilst playing a game, hence many participants find it unnatural which can potentially affect the gameplay experience. Furthermore, if the timing aspect of the game is integral to the game mechanic then such talking will affect this.

Similar to using the observation method, the session needs to be designed carefully so that the participants would feel comfortable, and to reduce potential biases as explained earlier. It is common practice to record (both audio and video) the user test sessions so that the data can be later reviewed and analysed to draw a stronger conclusion.

Although think-aloud protocol was not used during the gameplay sessions of the thesis studies as it can potentially affect players' physiological measure. However, study S2 looks at players' natural verbal comments as part of mixed methods to study their social interaction. Likewise, in the study presented in Chapter 6, players were asked to watch their gameplay video to comment on their actions and experience, much like using think-aloud protocol while reviewing their gameplay video in retrospect.

2.2.3 Heuristic Evaluation

Heuristic evaluation promises to provide a formal and accessible usability evaluation method, which can be used even before any code has been written. The method is rooted in usability research and should be performed by experts. Either they play a test game and compare the interactions against a chosen heuristic set, or observe players interacting with a game to compare players' behaviour against a chosen set of heuristics. The aim is typically to provide feedback if the game breaks any heuristic or design guidelines, and as such the evaluation may also suggest some solutions to the identified issues.

There are a number of different heuristic sets created for video games evaluation, such as (Desurvire & Wiberg, 2009; Federoff, 2002; Nielsen, 1994; Pinelle, Wong, & Stach, 2008a) . Although heuristic evaluation promises to be a low-cost usability evaluation method, it suffers significantly with problems concerning of evaluators' subjective interpretation (White, Mirza-Babaei, McAllister, & Good, 2011). To answer this limitation, researchers have aimed to develop a more specific set of heuristic. For example, to fit a certain game platform (e.g. for mobile game (Korhonen & Koivisto, 2006)), genre (e.g. for real-time strategy games (Sweetser & Wyeth, 2005)), genre weightings (Pinelle, Wong, & Stach, 2008b) or critic proofing approach (Livingston, Mandryk, & Stanley, 2010) that takes into account a problem's frequency, impact, persistence and a game's genre.

Heuristic evaluation was not directly used in this thesis, but they were employed to classify issues uncovered from different approaches in study S1, categories obtained from PLAY heuristic set by Desurvire & Wiberg (2009).

2.2.4 Questionnaires

Questionnaires are a frequently used user research method as they allow a relatively easy and fast collection of large amounts of data. They are often generalisable, convenient, and amenable to rapid statistical analysis, which is valuable to providing an insight into data collected from

other user research methods. In video games evaluation they often feature a Likert-scale (Likert, 1932) and are used in conjunction to gameplay observation; for example, players are asked to answer a questionnaire immediately after the gameplay session (before any discussion to reduce bias) in order to capture their experience while it is still fresh in their mind. It is also possible to ask players to fill in a questionnaire during the game play, for example when a natural break presents (e.g. end of level) to capture player experience in those intervals (such as in Kim et al. (2008)). Yet questionnaires only generate data when a question is asked, and interrupting gameplay to ask a question could be disruptive. If users are provided with questionnaires after the gameplay, rather than continuously throughout its course, their responses would reflect the finished experience and therefore important issues may not be identified (Mandryk, Atkins, & Inkpen, 2006).

Questionnaires have been used to collect a variety of information related to players or/and their game player experience, for example to understand player demographics, player type (e.g. (Andreasen & Downey, 2003), this will be revisited in Chapter 3), evaluation of player experience e.g. game experience questionnaire (GEQ) (Poels et al., 2007), immersion e.g. (Jennett et al., 2008), emotion (Lang, 1995) or engagement e.g. (Brockmyer, Fox, Curtiss, & McBroom, 2009). While questionnaires are widely used in understanding players and their gameplay experience, they can often lack in providing deep information (in comparison to e.g. interview) and can be less informing if used on a small number of players and not carefully designed.

Studies presented in this thesis also utilised questionnaires. In all studies participants were asked to complete a short questionnaire about their demographics (see Appendix 5). In study S2 players were ask to complete an online questionnaire (Andreasen & Downey, 2003) prior to the gameplay session in order to identify their player type. In the study presented in Chapter 6, after each game condition, participants completed four questionnaires (PANAS (Watson, Clark, & Tellegen, 1988), SAM (Bradley & Lang, 1994), SUS (Brooke, 1996), a Likert questionnaire on the game features - see Appendix 6). Additionally following completion of all conditions, participants completed a final rating (see Appendix 7) soliciting their opinions of the three game prototypes.

2.2.5 Interviews

As discussed above questionnaires are somewhat limited in data collections usually resulting in broad but not deep enough insight. One problem is participants tend to answer questions that are asked directly, leaving open-ended questions usually unanswered. Interviews are one possible approach that allows a deeper understanding of participants' needs.

Interviews are usually conducted with fewer participants but allow deeper discussion, especially on the topics (or questions) that matter most to researchers or participants. However, the quality of data collected through interview depends on an interviewer's skills and experience to not only ask the right questions but also on how to ask them (Lazar, Feng, & Hochheiser, 2010). Moreover, similar to other user research methods choosing participants is critical in order to gather useful data. Interviews are usually conducted with experts who can provide an insight into the project (e.g. the study reported in Chapter 4 utilised interviews to get professional game developers' insights on the BioSt early prototypes) or with participants to get information about their experience (e.g. the study reported in Chapter 6).

Depending on the study, an interviewer can take different approaches to asking questions. Usually there are some specific questions that need to be asked, but interviews can also be extremely flexible, for example in re-ordering questions or asking follow up questions. However, similar to other self-reporting techniques in video games evaluation, when participants provide information after the play test, they will often not recall motives for their actions, potentially leading to post rationalisation. These weaknesses can be addressed to an extent by recording the game video and replaying sections of interest in order to facilitate recall (e.g. (Gow, Cairns, Colton, Miller, & Baumgarten, 2010)). However, similar to other video analysis techniques, this is highly time consuming and would be less practical for a longer play test session. Chapter 3 details a technique for effective post play test interview, utilising changes in player's physiological measure of arousal.

2.2.6 Metrics

HCI researchers have been using automated data collection approaches to collect vast amounts of data, for example on user interaction with a test product. Lazar et al. (2010) summarised these approaches on a spectrum of "ease of use" (collecting data using existing infrastructure) and "flexibility" (customised or instrumented tool to collect specific data). The balance is between using existing tools than can be easily applied for research (e.g. website access log), or to create custom-built tools which provide more flexibility, higher capabilities and are optimised to answer specific research questions but are more difficult to conduct.

GURs leverage the use of metrics to better understand players' actions and behaviour by collecting large amount of data, for example from player interactions or player position (progress) in the game world. One of the benefits of metrics is that it enables the analysis of long-term player behaviour (Drachen & Canossa, 2009a), which is particularly useful for ensuring play balance (Hullett, Nagappan, Schuh, & Hopson, 2011), especially that balancing issues may not easily detect by classic user testing approaches. Game developers usually need to include codes that allow logging data (more on the "flexibility" side of the spectrum discussed

above), which means game developers need to work closely with GUR professionals in order to decide which metric to collect and program relevant code for. This can add an extra load to an 'already' complicated game development cycle. However, if implemented successfully they can detract this load in a long run. There are also attempts to invent approaches for collecting metrics without using special code in the game. For example: Marczak, van Vught, Nacke, & Schott (2012) detailed a visual and audio analysis technique that can collect some of these metrics by analysing game output.

Similar to other user research methods, it is preferable for game developers to collect behavioural data as soon as possible. However, these data are usually collected toward the end, where there is a playable version of a game available and not always during the development cycle. Nevertheless, gathering data from released games may have other benefits. For example, Hullett et al. (2011) analysed long-term game metrics data to identify unused features in their test game that can be used to iterate game features for future releases. Another possible downside is to analyse quickly and accurately the large amount of data that is usually collected using metrics. To answer this several visualisation approaches have been introduced in order to assist GURs to better understand the data, such as Data Cracker (Medler, John, & Lane, 2011), PLATO (Wallner & Kriglstein, 2013a) and many other examples as discussed in Drachen, Canossa, & Sørensen (2013).

To summarise, researchers have used game metrics to uncover game design issues (e.g., (Moura, Seif El-Nasr, & Shaw, 2011) or to tweak level design (e.g., (Drachen & Canossa, 2009b)). Game metrics are powerful tools to monitor players' action (or behaviour) in the game world; the challenge is to uncover how they felt (their experience) or why they perform some actions (as discussed in (Nacke et al., 2009; (Drachen & Canossa, 2009a; (Lynn, 2013)). Canossa & Cheong (2011) also highlighted that the collected players' behaviour metrics is not necessarily an expression of the player's personality and intention. Hence, it is recommended to combine game metrics data with qualitative user research methods (Kim et al., 2008). This is one of the areas that the work in this thesis may eventually contribute to, given additional development. For example, Chapter 7 discusses a possible development of Biometric Storyboards in a framework to visualise player's physiological measure (potentially linking to their feeling) and map them to the metric data (their action) in each game segment.

Overall, to better understand players' behaviour and their gameplay experience, GUR professionals use a combination of user research resources (as well as physiological measurement, which will be explored in next section of this chapter).

2.3 Physiological Evaluation

Physiological measurements can provide a continuous recording of player's physiological state. This section looks into physiological measurements, their evaluation, advantages and challenges in more detail.

There are directions emerging in GUR approaches making use of physiological data, still to be explored in terms of their actual usefulness in games development. Psychophysiology is a research domain that measures body signals to understand what mental processes are connected to measured body signals. Ravaja (2004) stated that the use of psychophysiology is a powerful research tool when examining communication, media, and media interfaces. This can also apply to video games as they share similar psychological phenomena such as attention, emotion and arousal that are of central importance in the research of media. These psychological phenomena have psychophysiological components, therefore physiological measures may provide important information on emotion that is complementary, or even contradictory, to that provided by self-report or observation. Ravaja (2004) mentioned physiological measures can also be regarded as more objective compared to self-report; they may also provide information on player's feelings that, for some reason (e.g., subtle nature of the responses, repression), are not available to players' conscious awareness. Thus physiological-based methods (also called biometrics) are now being more integrated in game user research than before (Drachen, Nacke, Yannakakis, & Pedersen, 2010a; Mirza-Babaei, Long, Foley, & McAllister, 2011) and game development in the game industry, for example (Zammitto, 2011; Ambinder, 2011).

There is a growing interest in HCI more generally to explore the potential of physiological measurements to evaluate user experience, cognitive, motivational, and emotional responses (Fairclough, 2009), as well as an awareness of the challenges in using physiological measurements.

Hence careful monitor and control study design and play environment is needed when physiological measures are used as part of a mixed methods. Fairclough (2011) discussed his points to adapt physiological measurements for the evaluation of gameplay experiences. One of the advantages of physiological data to subjective methods is that they provide continuous monitoring of behaviour without the need to interrupt the player, however they are more sensitive than observing gameplay behaviour and can potentially be intrusive as the sensors need to be attached to the player's body. In addition to a signal-to-noise issue, one challenge arises from the 'many-to-one' relationship between psychological processing and physiological response, that allows for physiological measures to be linked to a number of psychological structures (Cacioppo, Tassinary, & Berntson, 2007). For example, a player could be emotionally aroused, not because of specific in-game elements but as a response to an external activity,

physical activity, anticipation, novelty, habituation or as a result of something not otherwise observed.

Moreover, if comparing different conditions, carefully controlled experiment design is essential. As such, recruited participants should have a similar level of experience and too many variables should not be changed at the same time. These types of experiments usually focus on some particular aspect of the player experience and select relevant physiological measure to answer the particular questions. For example, if the focus of study were to capture the level of mental workload then measures of cognition would be more suitable.

In order to provide context to sensitive physiological measures, they are usually used in conjunction with other user research methods. So far this chapter has detailed some of the most common methods for GUR, thus this section focuses on introducing physiological measurements and discusses their contributions and limitations in GUR. It briefly looks at how the human body reacts to events, how these reactions can be captured (sensors) and analysed, and how game researchers utilise them to get a better understanding of players and their gameplay experience.

As mentioned earlier psychological phenomena (such as imagination, anticipation or perception) have psychophysiological components that can effect players affective state during gameplay, that lead to them making decisions (or actions) which itself can lead to a new set of imagination, anticipation or perception (Figure 2-2). Therefore physiological measurements can provide information on changes in players' affective state.

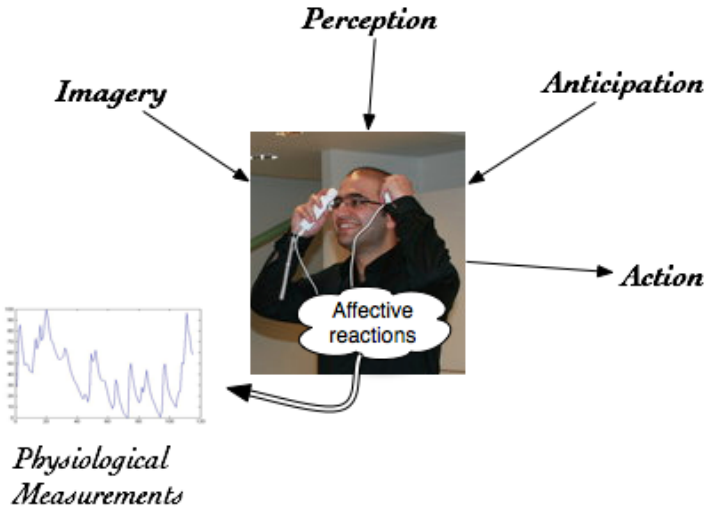


Figure 2-2 Three components are involved in recording physiological metrics: external physical activity, internal emotional activity and internal cognitive activity (Nacke, 2013)

Nacke (2013) sees this distinction between emotional triggers as especially relevant when analysing psychophysiological reactions together with game events: "Only through the

use of game logs that pinpoint exactly what game events were happening when, are we are able to contextualize physiological reactions of players (p. 589)". This thesis shares the similar view, by triangulating player comments and game events with physiological data.

As discussed before, one of the advantages of using physiological measurement is that in most cases they are recorded unconsciously (hard to fake), which makes them more objective than self-reporting approaches. Also most of these measurements can be recorded continuously, meaning we do not interrupt players and break the gameplay to get an idea of how they are experiencing the game.

2.3.1 Physiological Measurements

The human nervous system (that has an essential role in the control of behaviour) is split into two parts: one is the central nervous system (CNS), which contains the majority of the nervous system and consists of the brain and the spinal cord. The other is the peripheral nervous system (PNS or PeNS), which contains the nerves and nerve cells outside of the brain and spinal cord. The main task of the PNS is to connect the CNS to the limbs and organs. As the PNS transmits our physical sensations, and unlike the CNS is not protected by the skull and bone of the spine, it makes it easier to access measurements via our skin. The PNS itself is also divided into the somatic nervous system (SoNS), which regulates voluntary bodily activity, and the autonomic nervous system (ANS), which takes care of our unconscious responses and controls internal organ functions. Controlling unconscious responses make the ANS more suitable for physiological evaluation. The ANS is also divided into two subsystems: sympathetic nervous system (SNS), which is to mobilise the body's nervous system fight-or-flight response, and the parasympathetic nervous system (PSNS), which is responsible for the stimulation of rest-and-digest or feed-and-breed activities.

Sensor technologies enable researchers to use physiological measurements for testing or quantifying a user's feelings. Depending on the research question (dimensions of feeling to explore) there are various measures that can be taken using different physiological sensors. Common physiological measures in game research include skin conductance (SC), electromyography muscle measures (EMG), Electroencephalography (EEG), skin temperature, respiration rate and electrocardiography (ECG), where several measurements can be computed from ECG such as interbeat intervals (IBIs), heart rate (HR), heart rate variability (HRV) (Kivikangas et al., 2011a).

This thesis utilised galvanic skin response (GSR) computed from SC and facial EMG, with the aim to cover both dimensions of users feelings (see Figure 2-3), hence the next two sections (2.3.1.1 and 2.3.1.2) explore these two measures in more detail and the reasons they were chosen (despite other possible measurements) to answer the thesis's research questions.

Common approaches distinguish physiological analysis on a temporal dimension: Studying phasic psychophysiological and behavioural responses at game events (points in time) (e.g.(Ravaja, Turpeinen, Saari, Puttonen, & Keltikangas-Järvinen, 2008)) and studying tonic responses to variations of in-game variables (time span) (e.g.(Mandryk & Atkins, 2007)).

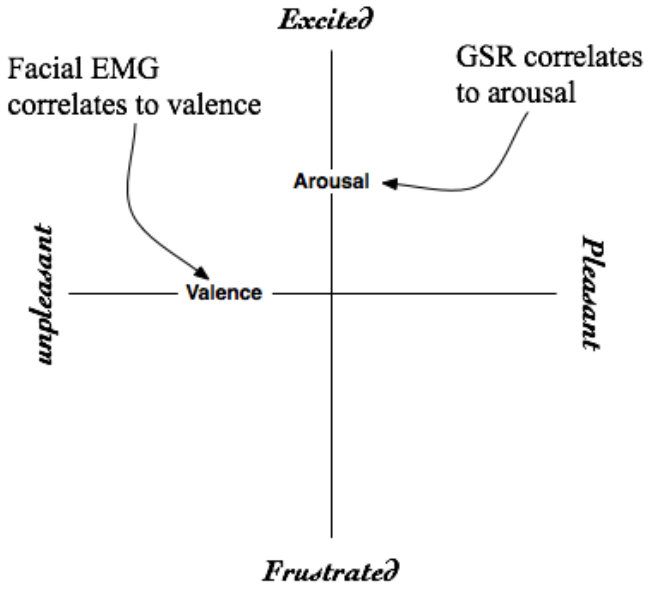


Figure 2-3 Russell, Weiss, & Mendelsohn, (1989) model of arousal and valence. The relationship between arousal/valance and physiological signals has since been applied to measure player experience.

Explanations of other mentioned physiological measures are not covered here, as this was not the focus of this thesis. However there are plenty of resources available; for example (Cacioppo et al., 2007) explored these measures in detail. In terms of the applications of these measures, section 2.4 provides some examples of how studies used them.

2.3.1.1 Galvanic Skin Response (GSR)

Arousal is commonly measured using SC (Lang, 1995), also known as galvanic skin response (GSR which is used in studies conducted for this thesis - when SC is measured as a direct response to a stimulus) or electro dermal activity (EDA - when SC is measured over time) and depends on how it is computed (Boucsein, 1992).The conductance of the skin is directly related to the production of sweat in the eccrine sweat glands. In fact, subjects do not even have to be sweating to see a difference in GSR because the eccrine sweat glands act as variable resistors on the surface. As sweat rises in a particular gland, the resistance of that gland decreases even though the sweat may not reach the surface of the skin (Stern, Ray, & Quigley, 2001). GSR has a linear correlate to arousal (Lang, 1995) and reflects both emotional responses as well as cognitive activity (Boucsein, 1992).

It is recommended to place GSR sensors to the fingers, palms or toes as there are more sweat glands in those areas, which make them more likely to react to changes in PNS. Although electrodermal activity can be measured from any of these sites, the values obtained are not necessarily comparable.

Because of relative ease of measurement and quantification combined with its sensitivity to psychological states and processes, GSR measures have been applied to a wide variety of questions ranging across examining attention, information processing and emotion (Cacioppo et al., 2007).

GSR has been closely linked with the psychological concepts of arousal and attention. Woodworth & Schlosberg (1954) supported this indexing relationship by noting that tonic GSR is generally low during sleep and high in activated states, such as rage or mental work. They also related phasic GSR to attention, noting that such responses are sensitive to stimulus novelty, intensity and significance. This thesis utilises phasic GSR since it focuses on the analysis and visualisation of player's arousal state and mapping this to in-game micro events.

There are two methods for measuring GSR: One, exosomatic: which relies on the passage of an external current across the skin. Second, endosomatic: which is recording the skins potential response without an external current. Most commercial physiological measurement kits use exosomatic methods for recording GSR, and this is the method of choice among researchers (Fowles, 1986).

The principle in the measurement of skin resistance or conductance is that of Ohm's law, which says skin resistance (R) is equal to the voltage (V) applied between two electrodes placed on the skin surface, divided by the current (I) being passed through the skin (R=V/I). Therefore, if the current is held constant then it is possible to measure the voltage between the electrodes (which will vary directly with skin resistance). Alternatively, if the voltage held constant, then the measure of the current flow would be skin conductance (which will vary directly with the reciprocal of skin resistance). Conductance is expressed in units of Siemens and measures of skin conductance are expressed in units of microSiemens (μS). Chapter 5 provides details on computing GSR measurements as utilised in the BioSt tool and study reported in Chapter 6.

For studies reported in Chapters 3 and 4, GSR data was gathered using the BIOPAC hardware system, sensors and software from BIOPAC Systems Inc. This was measured by using two passive SS3LA BIOPAC electrodes. The electrode pellets were filled with TD- 246 skin conductance electrode gel and attached to the ring and little fingers of the participant's left hand. Electrodes sites should be in a natural condition and not be cleaned by alcohol or abrasion. However, it is recommended for participants to wash their hand with a nonabrasive soap before

having the electrodes attached on to clean and dry skin. Room temperature, humidity and time of day are two environmental factors that should be controlled (Hot, Naveteur, Leconte, & Sequeira, 1999). Boucsein (1992) recommended a room temperature of 23 C.

2.3.1.2 Facial Electromyography (EMG)

As discussed earlier emotions can be interpreted in a two dimensional model: arousal and valence (Russell et al., 1989). While GSR can help us on uncovering feelings related to arousal, the final study of this thesis (Chapter 6) utilised measurement of facial muscles as a way to interpret valence. Hence this section looks into electromyography (measurement of muscle activity) and more specifically facial muscles.

To measure whether a muscle is active or not, an EMG electrode needs to be attached to the surface above a muscle to be able to sense the slightest activation (Lang, 1995). Therefore, depending on the placement of EMG sensors, most muscle activities can be measured. However, measurement of facial muscles is the most established for evaluating valence (Fridlund & Cacioppo, 1986). Especially, measurements of brow muscles (corrugator supercilli) and cheek muscles (zygomaticus major) to indicate positive or negative reactions to game events (Hazlett, 2008).

As EMG sensors measure signals released due to a muscle activity they need a reference for comparison. This reference sensor should be placed on an area without any muscle; for facial measurement it is common to attach the reference sensor to an ear lobe. Placing sensors on a participant's head can be intrusive and introduce movement artefacts. Also as facial muscles will be easily activated (for example by talking) participants need to be informed not to talk or move their head while signals are recorded, thus blocking any form of talk aloud protocol. Although these artefacts mean careful data interpretation, analysis of EMG signals is not complicated.

Chapter 5 explains EMG analysis for BioSt tool in detail, where NeXuS-10 MKII device and sensors were used to record GSR and facial EMG for the BioSt tool. For Zygomaticus major (smiling) and corrugator supercilii (frowning) facial muscle activity were measured using passive EMG sensors on a player's cheek, brow, and ear lobule (for ground sensor). Recording software was a custom C++ application using the NeXuS SDK to collect raw data from the device and display the recording timestamp on the computer screen.

As discussed earlier, physiological measures are often applied in conjunction with other user research methods to provide context to these sensitive measurements. So far this chapter provides the introduction to physiological measurements and discussions on their contributions

and limitations. The next section looks at how these measures are used in combination with other user research tools.

2.4 Mixed Methods

Applying mixed methods to evaluate user experience in video games has been recognised as a favourable approach that allows the effect of natural limitations associated with different user research methods to be minimised (Ijsselsteijn, de Kort, & Poels, 2008; Pagulayan & Steury, 2004). Therefore a triangulation of methods is applied in most of the GUR studies and this section looks at some examples of these.

This section looks at details of methodological studies where their main goal was to introduce a new combination of mixed methods for a better evaluation of player experience. As well as studies where their main goal was to deal with reporting issues, such as different visualisation techniques and reporting formats.

A number of industry-standard usability approaches use traditional methods in combination; for example, RITE (Rapid Iterative Testing and Evaluation) (Medlock, Wixon, & Terrano, 2002) which employs observation and think-aloud techniques with the addition of an attending software engineer to rapidly alter the design, based on the findings of the usability testing. Changes can be made after observing as few as one participant, with altered designs subsequently tested on the remaining participants. Similarly, the TRUE (Tracking real-time user experience) instrumentation methodology uses multiple methods to gather behavioural, attitudinal and contextual data to have a better understanding of player experience in addition to observed gameplay (Kim et al., 2008). In TRUE approach, user initiated events are automatically recorded along with contextual data; for example a crash event in a racing game is captured along with the difficulty level and game conditions. With the aim to provide a rich source of behavioural data for analysis, which can then be combined with attitudinal data, gathered either from a post-session questionnaire or during the game. The single player mode of Halo 2 was used to demonstrate an application of the method, in which the user was prompted every three minutes to give feedback on the game. While this technically allows player responses to be captured during the gameplay, and can be used to improve video games usability, breaking the gameplay can potentially impact on the game experience; hence this approach may be less suitable for studies focusing on the player experience.

Since determining how player experience changes during a gameplay is valuable information for a game developer, different approaches have adopted pausing the game to obtain feedback from a player to capture their feelings over the gameplay session. This can be based on pre-determined time intervals (Kim et al., 2008; Drachen, Nacke, Yannakakis, & Pedersen, 2010b)

or on sets of target game events (van Reekum et al., 2004). However, both approaches break the gameplay continuity, and may not accurately represent moments in which the player experience varies.

As discussed earlier questionnaires are frequently used to assess player experience. While the forced choice in questionnaires may provide useful data on player experience, they also lack on a deeper exploration of issues raised during the study, which is essential to further explore underling reasons behind each identified issues. Thus, questionnaire based studies are usually followed by interview sessions to allow a deeper discussion on areas of interest.

Overall, if we interrupt players to ask about their experience we break their gameplay and potentially effect their experience, whereas if we wait until the end of the gameplay players provide feedback on the final (complete) experience, instead of continuous during the gameplay. Many approaches are proposed to tackle this.

One possible approach would be taking cues from players' responses to gameplay (without interrupting them) to identify moments in the gameplay for further discussion with the player. For example, direct observation of players' facial expressions or body language has been used to investigate how players' feeling change during the gameplay (Lazzaro & Keeker, 2004). Given that this may reveal details of player experience, direct observation can be considered as an appropriate identification mechanism for moments of significant change in players' feeling during gameplay.

Lazzaro & Keeker (2004) discussed a study where direct observation has been used to identify player emotions by analysing players' facial expressions, body language and verbal comments. The study collected observation data from 40 participants playing their preferred games and describing their thoughts across the gameplay experience. The data was sorted into similarity groups that were used to identify four paths for achieving positive emotion in games and it was concluded that direct observation could reveal details about player emotions (Lazzaro & Keeker, 2004). The study by Barendregt & Bekker (2006) is another example of using behavioural observation to identify usability and fun problems. In this study a coding scheme was developed based on the DEVAN (Vermeeren & Bouwmeester, 2002) method, applied to detect usability problems in task-based products for adults. For example, scheme coded behaviours such as 'bored' or 'puzzled' and the proportion of agreement between two observers was used to validate the reliability of the coding scheme. However, the scheme does not suggest which coded behaviours are indications of usability problems in games and what can be fixed in the game.

Fabricatore, Nussbaum, & Rosas (2002) conducted a study to investigate the key concepts of playability in video games. The study used grounded theory[1] for the analysis of comments and interview data with guiding questions that were used to focus the development of categorisations to the relevant topic of playability being explored. The transcribed interview data of 53 participants was broken down into sorted categories. Similarly, grounded theory was used by Brown & Cairns (2004) for the analysis of players' interviews to develop a concept of game immersion .

On the other hand, in contrast to grounded theory, some researchers take a more exploratory approach. For example, Poels et al. (2007) identified themes and comments of interest quoted to propose loose categorisations that maintained the richness and variability of the responses. An expert panel of game researchers then reviewed these categories from which a final categorisation for game experience was proposed.

Although approaches using classic user test methods can potentially provide a rich source of data, understanding player experience require measures that can take into account how player engagement can change during gameplay events (Bernhaupt, 2010). Physiological measurements are a powerful tool that enables researchers to collect continuous body reaction to events. Hence they are becoming a popular measure in studying human behaviour.

Traditionally physiological data is been used to measure physical activities. For example, Segal (1991) measured heart rate, blood pressure, and oxygen consumption while their participants (32 males and females aged 16 to 25 years) played a video game and compared them with measurements made in a standing but inactive position. Their results showed that playing the video game significantly increased heart rate, systolic and diastolic blood pressure, and oxygen consumption. However, physiological measures have also been used to differentiate between human emotions such as anger, grief and sadness (Ekman, Levenson, & Friesen, 1983). So it is possible to argue the change in players' heart rate or oxygen consumption in Segal (1991) study were not due to change in physical activity but as a result of change in players' emotional state.

Researchers in different fields have been using physiological measures to get a better understanding of user behaviour. For example, in product design Laparra-Hernández, Belda-Lois, & Medina (2009) used EMG activities from the zygomaticus major and corrugator supercilii muscle regions and GSR for evaluating user perception. Their study suggests that physiological measurements would contribute to the understanding of user perceptions by incorporating measurements that do not involve conscious processes.

[1] Grounded theory is an approach to analysing qualitative data which allows emerging theories to be formed from the data rather than working from formulated hypotheses (Glaser & Strauss, 2009).

Physiological metrics have also been used in interactive systems. For example, Ward & Marsden (2003) used GSR and cardiovascular measures to examine user response to well and ill-designed web pages. Hazlett & Benedek (2007) used facial EMG as feedback in the software design process; they demonstrate the usefulness of measuring involuntary emotional reactions at key product purchase evaluation stages: first-impression (aesthetic) and during use (interaction). There are also attempts to use GSR as an index of cognitive load, for example, the study by Shi, Ruiz, Taib, Choi, & Chen (2007) showed potential results in explaining the peaks in the GSR data, which they found to correlate with user cognitive load in sub-task user events.

Physiological measurements are a powerful tool that enables researchers to collect players' reactions to game events without interrupting gameplay (unconscious and continuous). For this reason there is an increase interest to use these measures in studying player experience. Researchers have proposed methodologies to integrate physiological measures in mixed methods:

Chanel, Rebetez, Bétrancourt, & Pun (2008) proposed an approach based on emotion recognition to maintain engagement of players in a game by modulating the game difficulty. In their study both physiological and self-report analysis lead to the conclusion that playing at different levels gave rise to different emotional states and that playing at the same level of difficulty several times elicits boredom. Similarly, by utilising the facial EMG Yun, Shastri, Pavlidis, & Deng (2009) presented a methodology to improve user's experience in computer games by automatically adjusting the level of game difficulty. The measurements are based on the assumption that the players' performance during the game-playing session alters blood flow in the supraorbital region, which is an indirect measurement of increased mental activities.

Studies of the application of physiological measures and evaluation to gameplay experience have had success, for example, Hazlett (2008) described the use of facial EMG as a measure of positive and negative emotional valence during an interactive experience. His study found that the corrugator EMG could measure negative valence during high intensity interactive play in spite of the confounding factor of mental effort. Similarly, in finding relationships between GSR and reported arousal (Mandryk & Atkins, 2007).

In series of experiments (Mandryk et al., 2006; Mandryk & Atkins, 2007) Mandryk et al. explored how the physiological measurements (GSR, Facial EMG and HR) responded to different gameplay conditions. For each physiological response, tonic measures for the mean, peak, min and standard deviation were calculated and compared to statistical analyses of the questionnaire results. Their results showed that mean GSR was higher when playing against a friend and this correlated to subjective self-reports through questionnaires related to arousal level. However, as discussed by the authors, by taking measures of the whole experience, it was

not possible to know if higher mean in GSR was due to a raise in the tonic level or that there were more phasic responses (peaks) across the game session. Moreover, GSR responses were analysed for overall goal of the game, hence limited in identifying specific gameplay events that caused higher GSR responses. Nevertheless, these studies show that GSR can be a suitable measure for responses to game events, and suggest that using phasic measures is needed for a more accurate study on the effect of specific game events in the overall player experience. Both points have informed the focus and scope of the research presented in this thesis.

Nacke & Lindley (2009) created a real-time emotional profile (flow and immersion) of gameplay. They measured electroencephalography, electrocardiography, electromyography, galvanic skin response and eye tracking responses. They have also collected questionnaire data after each play session. Their results demonstrate a correlation between subjective and objective indicators of gameplay experience that shows the potential for providing real-time emotional profiles of gameplay that may be correlated with self-reported subjective descriptions. Hence, using a response profile for a set of physiological variables enables researchers to go into more details with their analysis and allows for a better correlation between response profile and psychological event. Livingston, Nacke, & Mandryk (2011) utilised physiological measurements to show changes in player experience over the course of their study to explore the effects of reading positive or negative game reviews on player experience.

Kivikangas, Nacke, & Ravaja (2011b) designed an analysis tool for the study of physiological responses to emotional expressions. Their tool creates time-framed video clips of pre-defined game events and presents them after an experimental session to the participant for recreating their memory of the experience at the event point in the game. The physiological measurements (GSR, ECG, facial EMG, and EEG) were recorded during the entire experiment, only to get additional data into participants' subjective gameplay experience. The tool aims to provide a triangulation between game events, phasic physiological responses, and self-report measures without interrupting the gaming activity. However, as the author's also noted, one challenge would be to pinpoint a specific event in a series of game tasks for self-reporting without relying on interrupting gameplay and administering a questionnaire or interviewing about the event retrospectively. The approach detailed in study S1 (Chapter 3) of this thesis aims to facilitate this challenge by using peaks in GSR measurements to pinpoint a specific event for structuring post-session interviews.

Although the study of physiological responses to emotional expressions in video games has become more popular over the last decade, but physiological-based evaluations are less common in the game industry. Apart from few cases within big publishers (i.e. Zammitto, 2011; Ambinder, 2011), there are not many reports if smaller to midsize studies have successfully

applied these measures. However, some independent GUR studios, such as Player Research[2], Immersyve[3] and Bunnyfoot[4] are promoting the benefits of various physiological metrics for player experience evaluation. Overall the barriers toward using physiological measurements, (such as that GUR personnel need to be trained extensively in interpreting and correlating physiological metrics with other measures), seem out of scope for small to midsize development companies. One step towards making these measures more available for larger GUR community is to solve the problem of 'easy interpretability'. As discussed earlier changes in the physiological signals can be responses to external activity or can be in anticipation of something not otherwise observed. Moreover, specific types of measurement of different responses (such as GSR, EMG, ECG and EEG) are not trustworthy signs of well-characterised feeling. The often described many-to-one relation between psychological processing and physiological response (Cacioppo et al., 2007) allows for physiological measures to be linked to a number of psychological structures (for example; attention, emotion, information processing). Ambinder (2011) emphasised that some responses or measurements are difficult to correlate with something specific that happened in the game.

Therefore, researchers using physiological measures often look for correlations between the collected physiological measures and the self-report measures as a means of validating the quantitative values captured. For example: Mandryk (2008) GSR and questionnaire; Nacke (2009) EMG, EEG, and GEQ; Lin, Omata, Hu, & Imamiya (2005) HR, GSR and questionnaire. One of the advantages of using physiological is that they provide continuous measures of the play experience, however mapping this to one-point qualitative feedback from a single questionnaire can be challenging (Mandryk et al., 2006).

This thesis is not looking into these types of correlations but instead into using self-reports (interviews in particular) to get the participants to explain the change in their measurements. This is based on physiological measures as a tool to provide continuous data on players feeling and self-report to explain changes in feeling. This would provide an easy analysis of physiological measures and also tackle the limitations of self-report, to get direct detailed feedback from players in comparison to questionnaires, which reflect on the finished experience. As such, the use of GSR as a way of identifying moments in the gameplay of higher impact to the player experience to be discussed in a post- session interview (Studies S1 and S2 in chapter 3). This could be effective in highlighting periods of the gameplay that may not be identifiable

[2] www.playerresearch.com

[3] www.immersyve.com

[4] www.bunnyfoot.com

using traditional observational methods, hence enhance user research approaches for GUR studies.

So far this chapter has discussed how user research method have been adapted and applied in combination to uncover usability and user experience issues in video games. However, one other important area of GUR is reporting. Using the correct user research methods and analysing data thoroughly would not improve a game if the findings are not communicated well, are not believable, or are not acted on. Thus, visualisation is a continuously growing area, with research efforts expanding into many different domains. Visualisation tools address the challenge of analysing and presenting overwhelming amounts of data. Different visualisation tools have been developed to analyse player behaviour data, such as Data Cracker (Medler et al., 2011), Lithium (Hoobler, Humphreys, & Agrawala, 2004), PLATO (Wallner & Kriglstein, 2013a), PlayerViz (Dixit & Youngblood, 2008) or Playtracer (Andersen, Liu, Apter, Boucher-Genesse, & Popović, 2010). Each used various technique for visualising gameplay data, these techniques vary from charts and graphs to diagrams, heatmaps and node-link approaches.

For example Data Cracker (Medler et al., 2011) was built for monitoring gameplay behaviour in Dead Space 2. The tool's features include; a summary graph to show overview values of the collected gameplay data, a timeline that displays the range of days for analysed data, and a main graph that displays data related to unique/total users being monitored, kill/death ratios, number of rounds played and won, experience points gained by players, weapon statistics, and objective completion rates.

The Lithium system (Hoobler et al., 2004) aimed to enhance the observation of user interactions by visualising player distribution over time and areas of combat with features such as the path taken by a player to their present position, fire tracer and fields of view in multi-player FPS games. For an example of heatmaps visualisations; Drachen & Canossa (2009a) used a grid-based heatmap of player death locations and causes in *Tomb Raider: Underworld.* In order to provide further contextual information, they represented different data sets as layers on top of the game level map.

Wallner & Kriglstein (2012) developed a visualisation system (PLATO) that visualises gameplay as a node-link diagram, where a game's states are visualised as nodes with the node-size being proportional to the number of players arriving at that state and actions are depicted as directed edges. The system offers various functions such as searching, filtering, clustering, chart generation and time-dependent simulation. Playtracer (Andersen et al., 2010) is another example of node-link representation of player behavioural data where nodes correspond to game states and edges show player's movement between states. To avoid over complexity the system uses feature-based aggregation to simplify visualisation.

In a recent review on gameplay data visualisation approaches, Wallner & Kriglstein, (2013b) show that there is an increasing interest in visualisations because they can help developers to analyse large amounts of gameplay data and better understand player behaviour. Graphical representations make complex gameplay data more understandable, however, most of these techniques focus on player behaviour but do not address qualitative information on player experience, such as reasoning or feeling. Providing game developers with easy-to-understand, actionable and motivational player experience report is the ultimate goal of the Biometric Storyboards approach presented in this thesis.

2.5 Summary

Various user research methods have been applied to capture interactions and behaviour from players across the gameplay experience. Player behaviour is often studied based on subjective reporting and observational approaches, hence as such may not focus on the parts of higher emotional interest. The application of physiological measurements has shown potential to provide continuous monitoring of behaviour without the need to interrupt the player, however, it has not yet applied what elements this can inform on formative evaluation of player experience to enhance design decisions in game development cycles.

As discussed earlier, the application of physiological measurements is particularly valuable for providing continuous and unconscious data through the gameplay. This would enable researchers to perform analyses across gameplay events or segments. However, given the advances in capturing continuous quantitative data (e.g. physiological measurements), one challenge is to effectively capture qualitative feedback from players to associate with recorded quantitative data. This is essential to explain players' actions, their experience, as well as to provide context to changes in their physiological states.

Another major challenge is tying the results of physiological measures and one-point data from a questionnaire or interview in player experience reports together. This is because the data is radically different. Analysing large-scale or high-resolution data (such as physiological measurements) can be daunting and presenting results from these studies is often not straightforward. This is more important as one of the goals of GUR studies is to identify actionable results that will lead to game design improvements. Using visualisation techniques facilitates the understanding in these data. As discussed earlier, various visualisation techniques have been introduced in GUR. However, most of these techniques only focus on displaying large amounts of data captured directly as a result of gameplay without comprising qualitative or contextual data on players' emotional experience. Hence, most of these techniques are restricted to only answer 'what players do (their behaviour)', and fall short to address the player experience issues of 'why they did it', or 'how they felt'.

3 Mixed Methods Implementations

The main outcome of this thesis is Biometric Storyboards, a triangulation of user research methods (physiological measures to enhance observation and post-session interviews) to gather data on gameplay experience. Each of these underlying methods themselves is also developed as part of this Ph.D. research. This chapter first provides an overview of each sub-part. Then, the main body of this chapter contains a description of the thesis' initial studies, where I explored the contribution of physiological measurements in GUR. The knowledge gained from these studies helped me to iterate, develop and investigate the usefulness of these underlying methods in order to develop BioSt.

3.1 Triangulation of User Research Methods

Getting users to talk about and explain their experience is the easiest and most widely applied approach to understand ones experience (Albert & Tullis, 2013). However, self-reporting techniques are limited when conducting GUR in comparison to a usability evaluation on productivity applications ((Hazlett, 2008), p. 189). Firstly, self-reporting methods, such as questionnaires and interviews, are sampling methods, meaning that the players will be responding at a specific moment in time. If they fill out questionnaires during the game it interrupts the gameplay/flow and modifies their experience. However, if we wait until the end, then they may have forgotten what the real experience was like, or they may not remember correctly (Ravaja, 2004). Secondly, if we ask players to self-report, although these can potentially provide a rich source of data, we are relying on their awareness, recall, and cognitive filtering abilities to function before a response emerges, and covertly assess the experience.

This section explains three approaches, which are explored as part of this thesis to enhance self-reporting approaches for GUR. These approaches are also used to collect players' data for BioSt prototypes.

3.1.1 Player's Self-assessment Diagrams

One motivation for this research is an idea to find a way to capture players overall gameplay experience, without interruption from the user researcher. One approach tried as an exploratory task was to provide players with a blank graph paper and asked them (without interruption or prompting) to 'draw their experience' at the end of each level (which provides a natural break in the gameplay experience). The interest was to explore how much detail players could recall

immediately after the gameplay. The outcome was that these player-drawn diagrams (Figure 3-1) reflect on a player's overall experience of each level without many details. For example, these diagrams could identify the gameplay issues that the players may tell their friends about. Also these player experience diagrams seem to address the perception issue: what the player thinks happened, and what they can recall.

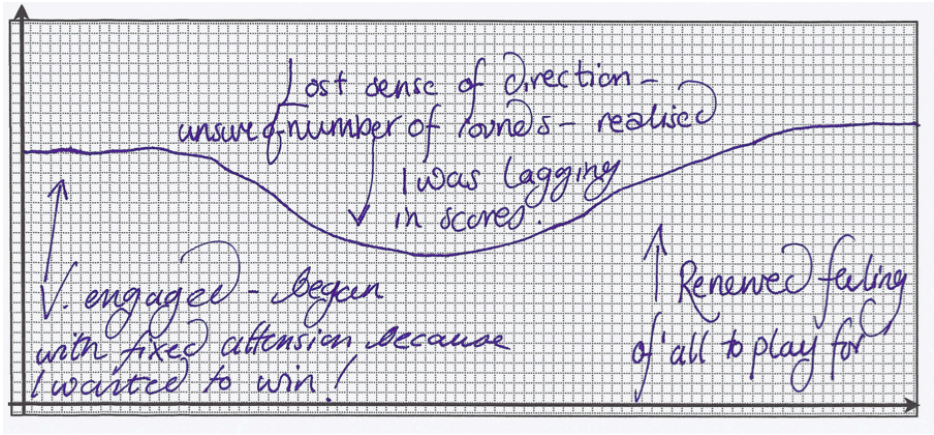


Figure 3-1 Example of a player's self-assessment diagram for 30 minutes of gameplay

On the other side, in the vast majority of cases, players could not accurately remember details of their gameplay experience, even after short game sessions. It seems that many players are only able to recall few details from the very beginning and the very end of that gameplay session. In most cases, the players draw a line graph which contains few peaks (or thoughts). In psychology this is known as the serial position effect (Feigenbaum & Simon, 1962). Broadly speaking, people tend to remember events at the start, the end, and perhaps one in the middle. Figure 3-2 shows an example of a player experience diagram that a player has drawn after just a 20-minute gameplay session.

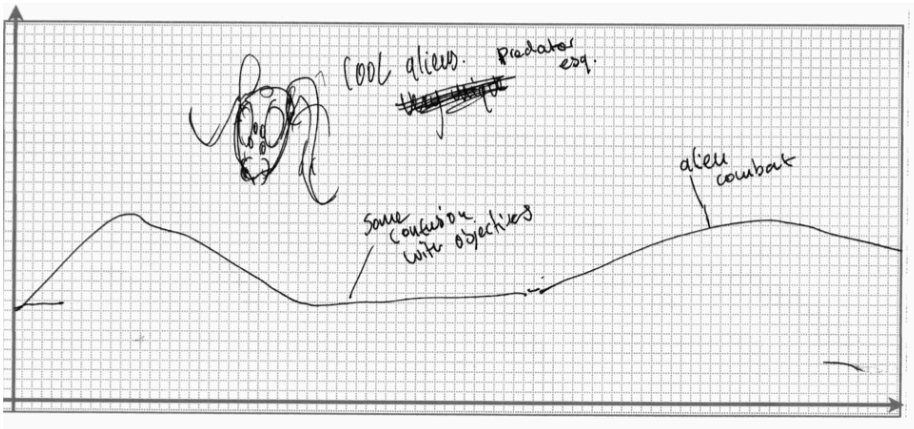


Figure 3-2 Example of a player's self-assessment diagram showing recalled events at the start and the end

These outcomes are all expected due to the limitations of self-reporting approaches as explained earlier. Later this chapter, section 3.3 shows how combining these players' drawings with

player's physiological responses and their post-session interview facilitates addressing this gap. These player's drawings are useful in a triangulation setting, providing extra evidence to support (or confirm) findings from other approaches (this is their contribution in creating BioSt).

3.1.2 Player's Physiological Arousal to Structure Post-session Interview and Coding Gameplay Events

Chapter 2 discussed that the use of physiological measures does not directly identify the feeling that a participant is experiencing. Generally, researchers using physiological approaches may find it difficult to match the obtained quantitative data to the participant's emotional experience during an experiment (van den Broek, Lisý, & Janssen, 2010). It is also possible to consider that player could be emotionally provoked, not because of specific in-game elements, but as a response to an external activity, anticipation, or something not otherwise observed. The often described 'many-to-one' relationship between psychological processing and physiological response allows for physiological measures to be linked to a number of psychological structures (Cacioppo, Tassinary, & Berntson, 2007).

Based on previous research, as discussed in Chapter 2, this thesis assumed a mapping of GSR arousal to player excitement (or frustration). The interest of this thesis is not to explore this mapping (it has been widely explored before). Neither to attempt to map the changes in a player's physiological measures to a particular emotion, instead using measures of the player's phasic physiological data purely to log 'micro-events' in the game. For example, a visual change (Figure 3-3) in the player's arousal level (peak in GSR measurement) are used to bookmark the player's gameplay video. These timestamps of 'micro-events' provide the structure for a post-session interview with the player.

Figure 3-3 Example of a peak in GSR signal

Thus, specific micro-events (identified by peaks in the monitored player's arousal levels) were noted during the playtest, constructing a log of times during gameplay in which players experienced a potentially meaningful degree of arousal. Micro-events were not analysed or interpreted at the logging stage, and at no time were individual participant's GSR measurements compared to other players. Instead, after the gameplay session, the video footage related to every logged micro-event was played back to players, who were asked to recall these specific moments and inform the user researcher of their thoughts. All logged micro-events were addressed in this manner, with usability and player experience issues determined by the players' interpretation of their physiological response.

The study S1 (section 3.2) demonstrates the value of this approach in the field of GUR and in generating BioSt. The results from this study show that using physiological arousal to drive post-session interviews would result in more meaningful insights into players' motivations and expectations.

As discussed in Chapter 2, observing players interacting with the game is one of the common approaches in GUR that provides a rich source of data. Physiological measures can shows the corresponding biological reaction from the player's body to game events. In addition to structuring post-session interviews based on the player's physiological arousal measurements, the techniques presented in the previous section form a framework for analysing the coupling between player behaviour (what they did) and feeling (how they felt).

The following sections of this chapter contain a description of two initial studies of this thesis, where I explored the contribution of explained approaches in GUR. The knowledge gained from these studies helped me to create, iterate, develop and investigate the usefulness of underlying methods in order to develop BioSt.

3.2 Study One: Using Physiological Arousal to Structure Post-session Interview[1]

This study aims to quantify the value of biometric method as an addition to classic user research methodologies, and their respective contributions to the production of formative feedback during the development of video games.

A series of user test sessions were performed in a dedicated GUR laboratory by three evaluators with professional and academic experience in conducting and analysing video game playtest

[1] The study presented in this chapter was published as a full paper at the DiGRA 2011 conference (see publication list P8). The thesis author was the leading author and researcher for the paper, designed and conducted the study, analysed and reported the results. The observation-based analysis was conducted by two HCI Undergraduate research assistants (S. Long and E. Foley) under the guidance of the thesis author. The statistical test was conducted by S. Hutton. This section is based on the published paper.

sessions. In this study, two usability and player experience evaluations were conducted separately on identical gameplay footage. A lightweight biometric-based experiment (as explained in section 3.1.2) was conducted 'live' during playtest sessions, and a typical observation-based approach was conducted using a dual-expert post-gameplay analysis on recorded video footage of the same session.

3.2.1 Data Collection and Setting

The games: Participants played the first two levels of 'Call of Duty: Modern Warfare 2' (MW2), developed by Infinity Ward in 2009 (Infinity Ward, 2009) and 'Haze', developed by Free Radical Design in 2008 (Free Radical Design, 2008). Both games are First Person Shooters (FPS) with Metacritic[2] review scores of 94% for MW2 and 55% for Haze. These titles were chosen to expose the study's methodology to games of differing quality, and therefore differing numbers and types of usability and user experience issues.

The playroom was equipped with a Sony PlayStation 3, a Sony 40" flat screen TV, a Sennheiser wireless microphone, a Sony Handycam video camera to capture the player's face and a BIOPAC system to capture physiological data. Participants were seated on a comfortable sofa positioned approximately two meters from the TV and the camera. The playroom was specifically designed and decorated to simulate an actual living room in order to reduce the impact of artificial experience.

Figure 3-4 GUR studio – playroom at Sussex University

The game footage, the camera recording the player's face, and the screen containing the physiological data (GSR) were synchronised into a single screen. This screen was digitally recorded and displayed on another screen in an isolated observation room, where the researcher were observing and controlling the gameplay session. Hence, participants were left alone during the playetest session and the researcher controlled and observed the session from the separate

[2] www.metacritic.com

observation room in order to reduce the experimenter effects. The digital recording also contained audio of the participant's comments and the game audio output from an attached microphone (Figure 3-5).

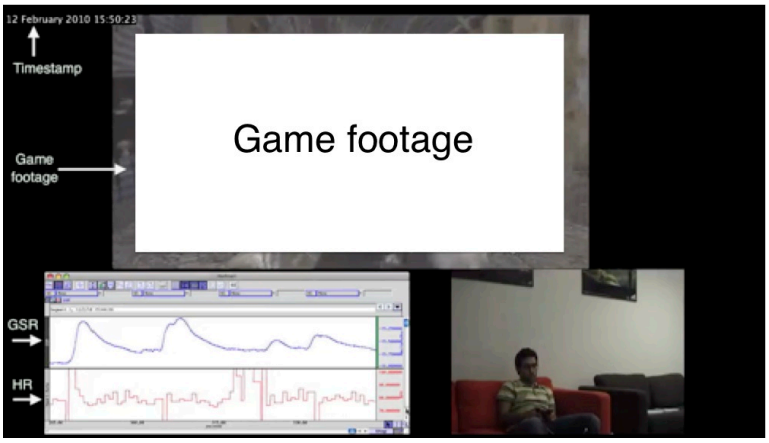


Figure 3-5 Example screenshot of the gameplay video

The participants: Potential participants from the University of Sussex filled out a background questionnaire, which was used to gather information on their previous video game experience, game preference, console exposure and personal statistics such as age. Participants were selected randomly and then recruited if fitted for the requirements of the study. Six male participants (aged 20 to 31) were recruited carefully from this list, ensuring they were casual PC or console gamers, with no previous experience of MW2 or Haze. All gave informed consent and there was no monetary incentive for their involvement. (Table 3-1)

	1	2	3	4	5	6
MWF2	P1	31	FPS, Strategic	PC, PS3	Everyday for 1 hour	Alone, online
	P2	21	FPS, Football	PC, PS3	2 hours every other day	With friends
	P3	23	Football, Racing, FPS	PC, PS3	Everyday, for 2 hours	With friends
HAZE	P4	20	Car racing, Diablo 3	PC	Everyday 2-4 hours	Always alone
	P5	31	FPS, Strategic	PC, PS3	Everyday for 1 hour	Alone, online
	P6	21	FPS, Football	PC, PS3	2 hours Every other day	With friends

Table 3-1 Participants' information: (1) ID, (2) Age, (3) Favourite games type, (4) Preferred platform, (5) Frequent gaming, (6) preferred play condition

Half of the participants were randomly assigned to play the first and second levels of MW2 in the normal difficulty mode. The other half played the first and second levels of Haze with the same difficulty settings. Both games were played on the Sony PlayStation 3 platform.

Setting: Upon arriving, the researcher welcomed each participant, explained the purpose of the study (without giving up too much details) and clarified that the focus is on evaluation the game and not on how well participants perform. After a brief description of the study procedure, they were then fitted with GSR sensors (Figure 3-6) and rested for few minutes to allow the researcher to baseline the signal.[3] The session took around 90 minutes (60 minutes of gameplay and 30 minutes of post-gameplay interview), depending on how fast they finished two levels. Overall participants provided over 6 hours of gameplay video for analysis.

GSR data were gathered using a BIOPAC hardware system (MP36), sensors and software from BIOPAC System Inc. The GSR was measured by using two passive SS3LA BAIOPAC electrodes (at 60 Hz). The electrode pellets were filled with TD-246 skin conductance electrode gel and attached to the ring and little fingers of the participant's left hand. The players were asked to relax for a few minutes so baseline measures could be recorded (Although this could be considered less problematic in this study as the focus was not on comparing GSR data among participants but to use the data to drive individual post-session interviews with the participants). Participants HR data was also captured using three pre-gelled leads with BIOPAC SS2L surface electrodes (at 50Hz), this was not the focus of this study and therefore was not included in the analysis.

Figure 3-6 GSR sensors attached to ring and little fingers

Post-gameplay: The post-session interview was conducted soon after they had finished the gameplay session so that the participants could remember most of their actions and thoughts. The interview was based on the selected events from the gameplay session, namely those events selected by monitoring changes in participant's arousal level (indicate by GSR measures). After each session, together with the participant, selected moments of their gameplay video were looked at and they described their feelings on those moments, but most importantly 'why they felt that way'.

Methodologies: The user testing data was subjected to analysis by two approaches, a biometric-based approach and an observation-based approach.

[3] Similar approach to study design and settings are followed for other studies reported in the thesis. As discussed the lab situation and the study setup are carefully designed to reduce potential experimenter effects and sampling biases, it is also assumed that the effects of these to be random.

Biometric and Post-session Interview: A lightweight biometric-based approach (as explained earlier) was applied, which does not attempt to analyse biometrics measures to interpret player emotional state, instead it uses measures of players' phasic physiological data purely to log 'micro-events'. These specific moments, identified manually by peaks in the monitored GSR levels for each player, were noted during the playtest, constructing a log of times during gameplay in which a usability or player experience issue may have been expressed. Micro-events were not analysed or interpreted at the logging stage; instead, after the gameplay session, the gameplay video footage related to every logged micro-event was played back to each player, who were asked to recall these specific moments and inform the experimenter of their thoughts.

All logged micro-events for each player were addressed in this manner, with player experience issues determined by the player's interpretation of their biological response. The data from post-session interviews helped the user researcher to identify usability or UX issues. The biometric-based approach used only the live feed of the gameplay to identify micro-events, and did not involve the review of video footage.

Observational Approach: In this observation-based approach, two evaluators analysed the same gameplay footage that was viewed and recorded during the study. Each evaluator watched and analysed all recorded gameplay videos individually, noting usability and UX issues in a 'double-expert' approach. Biometric readings were not taken into consideration in this post-gameplay analysis. Once each of the evaluators had completed the analysis of each gameplay video, their findings were collated and summarised with identical issues combined, providing a single list of findings. These issues are considered to be representative of those that could be found by the observation-based approach universally, both in content and quantity.

For example, a player became lost when attempting to follow a comrade as instructed in MW2, and ended up doubling back through several rooms in their confusion. The player's body language also reflected frustration. This behaviour indicated a usability issue with the location marker prompt to "follow", which was not visible from the player's original position.

Negative usability or user experience issues that were identified from player comments during gameplay were included in the results of this approach, reflecting the content of a typical observation-based methodology.

3.2.2 Results

The interest of this study is to identify the strengths, weaknesses and qualitative differences between the findings of the proposed structured post-session interviews using GSR and the results of a full observation user test study. By gaining an understanding into the contribution of GSR measures to GUR, this study aims to explore the integration of methods from across the

traditional qualitative/quantitative divide, as well as quantify the contribution of this approach as one of the foundational mix-methods to create BioSt.

In order to isolate the biometrically-determined findings of this approach, only negative usability or user experience issues indicated by the presence of players' biometric arousal were classed as findings. Any other issues noted by the player conversationally during the post-session interview were not included. These negative issues (circumstances evoking positive arousal were not included as usability or user experience findings) were then compared to the negative usability and user experience issues identified with the observational-based approach.

From the total of 89 issues found, 29 (32.6%) were identified by both approaches. Observation-based user testing established 34 issues (38.2%) that the biometric-based approach did not. Using the biometric-based approach, 26 issues were revealed that were not found in the observation-based user testing methods (29.2%). A total of 58 issues were identified in Haze, with the remaining 31 issues identified in MW2.

In order to gain a better understanding of the nature of the findings, issues were sorted into three categories (CAT1: Gameplay, CAT2: Emotion\Immersion and CAT3: Usability), allowing the strengths and weaknesses of the two approaches to be attributed to certain categories of usability or UX issue. These categories were obtained from (Desurvire & Wiberg, 2009) and are provided in Table 3-2.

CAT 1: Gameplay	
1.1 Enduring Play	1.2 Challenge
1.3 Strategy and Pace	1.4 Consistency in Game World
1.5 Variety of Players and Game Styles	1.6 Players Perception of Control
1.7 Goals	
CAT 2: Coolness/Entertainment/Humour/Emotional Immersion	
2.1 Emotional Connection	2.2 Coolness/ Entertainment
2.3 Humour	2.4 Immersion
CAT 3: Usability & Game Mechanics	
3.1 Documentation/Tutorial	3.2 Status and Score
3.3 Game Providing Feedback	3.4 Terminology
3.5 Burden On Player	3.6 Screen Layout
3.7 Navigation	3.8 Error Prevention
3.9 Game Story Immersion	

Table 3-2 Issue categories obtained from (Desurvire & Wiberg, 2009)

The categorised observations are shown in Figure 3-7.It is clear that the majority of CAT3 issues were revealed by the observation-based approach, whereas for CAT1 and CAT2 issues, the majority were revealed by the biometric-based approach. This observation is supported by

the fact that a chi-square test on the frequency of observations with categories (1, 2 & 3) and approaches (biometric-based, observation-based or both) as factors was highly significant – χ^2 (8, N=89) = 26.7, $p < .01$). This means there is a relationship between approaches and categories that goes beyond what would be expected by chance alone. Chi-square does not describe a relationship, instead it has to be interpreted from the data. It would seem sensible to conclude that the relationship here is CAT1 and CAT2 issues are better revealed by the biometrics-based approach than CAT3 issues. This will be discussed further here and in the discussion section.

Through categorisation of the results, it is clear that observation-based user testing revealed a significant number of those issues in CAT3, usability and game mechanics (90.4%). There was an overlap of 40.4% where those issues were also indicated by the biometric approach, with 9.6% of issues being identified only by the biometric approach in this category. Issues in CAT1 and CAT2, concerning players' feelings, immersion and gameplay experience, were more separated. The majority of issues found in these two categories were only indicated by the biometric approach (53.8% in CAT1 and 63.6% in CAT2). Observation-based user testing was less effective than the biometric approach (15.4% for CAT1 and 36.4% for CAT2). 30.8% of the CAT1 issues were found by both methods but there was no overlap in the CAT2 issues.

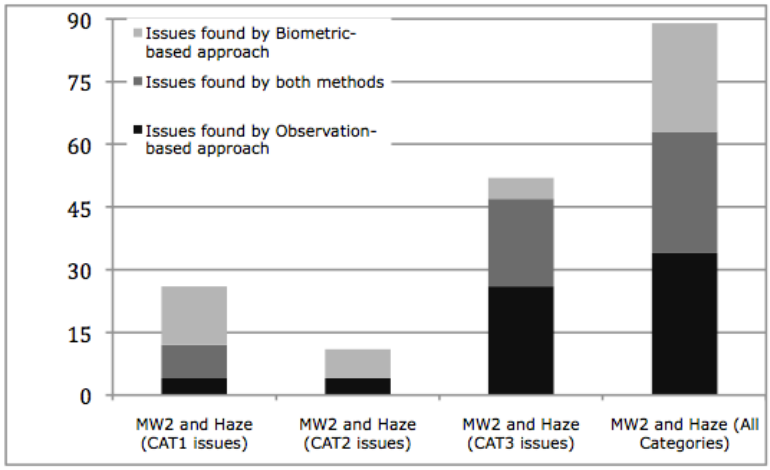


Figure 3-7 Comparison of number of issues

3.2.3 A Closer Look:

In order to get an in depth understanding of the quality of issues identified with the biometric-based approach, this section provides some examples of the results from the qualitative post-session interviews to explore the effectiveness of this approach to understand players' thoughts and feelings in the games. The following are selected examples of the findings:

P1 commented on the increased GSR measures when he entered a building from the street during the first level in MW2: "*You never get tired of this game, the game environment changes*

frequently, from streets to a building, then back to streets, then on the car". Indicating positive experience from variety in game environments.

In MW2, at the beginning of the second level players had to climb an ice rock with a pickaxe. It was something new in the game, which led to a constant increase in all the players' GSR measures, that could mean a high level of excitement. P2 commented on that event as: "*So interesting, awesome, I haven't seen anything like this before*".

In MW2, decreases in GSR readings were recorded for all the players when they finished climbing the ice rock before engaging enemies, at the beginning of the second level. The decrease in the measures could be assumed as they were getting bored, but in the post-gameplay interview they similarly commented on that moment as: "*Fun bit of relaxing after the heavily taxing moments of climbing the ice, it felt really good and I can't wait to see more of the game*" [P3]. Indicating positive experience from change in game pace.

Later on at the end of the level, players had to escape the enemies' camp with a snowmobile. A similar pattern of increase in all of the players' GSR was observed. They noted that event as one of the most exciting moments in the game. P1 said: "*The best thing was that riding it was not very difficult, since it is not a racing game. I was so excited, especially when I had to use my gun. It was realistic, I even had to reload my gun too*". Indicating positive experience from novelty in the game's actions.

Repeated increases in P3's signals were observed before he reloaded his gun in MW2. Later in the interview he commented: "*I've chosen a machinegun because it is so powerful but it takes time to reload it. It is annoying, but it is also realistic which I like. Every time I wanted to reload it I was afraid the enemy might attack me before the gun was reloaded*". Indicating positive experience from matching the game world with player's expectation.

P6's GSR signal rose when he was engaged with an enemy's car for the first time in Haze. It could be thought that this was because it was the first time he saw the enemy's vehicle in the game; but at the post-gameplay interview he mentioned: "*Since I was driving my buggy, I was expecting to be able to follow and shoot at the enemy's car, but the car went to a closed area like a warehouse. It was quite easy to kill them. It would be more challenging if I could follow and shoot at it while driving fast*". Indicating negative player experience as the game world did not match expectation of play.

In Haze, an increase in P5's GSR was noted when he was using the sniper rifle. One question was if he liked the weapon? In the post-gameplay interview he mentioned: "*In that event, I killed four enemies all together; all of them came to the game scenario from the same location, and it was very easy to kill them all. I was expecting a tougher experience for an FPS game*".

Indicating a negative experience from poor balance between player's abilities and challenges in game.

An increase in P5's GSR was observed while he was watching a cutscene in the middle of the first level of Haze; one thought could be he was enjoying this clip. In the post-gameplay interview he commented that: "*I was expecting a cutscene related to my mission, but this was very boring and I couldn't skip it*". Indicating negative experience from the game not matching player expectation and also usability issue with non-skip-able cutscene.

P4 explained the peaks in his GSR when he was driving the buggy later in the second level of Haze: "*Riding the buggy was quite fun at the beginning, but later on in the level I got so bored with it*". Indicating negative experience from lack of novelty in the game actions.

In Haze, there was a sharp increase in P6's GSR readings after a few minutes of starting a new level. It was hard to explain this increase as he did not mention anything during the game, but at the post-gameplay interview he said: "*I was feeling lost and not in control of the game*". Indicating a negative experience from a lack of guidelines leading to player's not feeling in control of the game.

Another example of rises in P5's GSR occurred while he was using the machinegun to kill an enemy solider. He mentioned earlier that he liked to use the machinegun, so perhaps that is why his GSR reading was rising. During the interview he described his feeling at that moment as: "*I was shooting at the enemy, hiding behind an aluminium board, so I thought I must have killed him, since he was hiding behind a very thin board. I was shooting with a powerful gun, yet he was alive*". Indicating negative experience cased by a lack in fulfilling player's expectation based on real world rules.

During different events in Haze, players described the reason for most of the increases in the GSR as that they were not sure if they were doing the right thing and a lack of feedback. For example P4 expressed: "*I was not sure if I could still drive my buggy or if it was broken. I've started driving it again, but was not sure if it was going to explode soon or not. Eventually, it did*". Or, P6 describes the increase in his GSR when a cutscene started in the middle of the first level of Haze as: "*I was not sure if I was walking in a right direction. I was lost in the jungle, so when the clip started I was relieved because I realised I was in the correct location*".

3.2.4 Discussion

Overall, observation-based user testing methods distinguished a greater number of issues, however, as literature has suggested, GSR provides only a measure of player arousal, which may not provide a representation of the full player emotion spectrum. Further research into the

number of usability issues uncovered using differing physiological sensors may suggest that specific sensors, or sensors used in combination, can reveal a yet greater number of issues.

Issue Quantity: The biometrics-based approach revealed a majority of gameplay issues, many of which were not identified through the observation-based method alone. The results demonstrate the important role that GSR, and potentially other types of biometric measurement, may play in conducting a thorough analysis of video games. Observational methods alone found the majority (90.4%) of issues in CAT3, but just 15.4% of those in CAT1. The addition of just one biometric measure increased the number of findings significantly, providing a valuable contribution to the formative analysis.

Issue Category: The results also indicate that there is a difference in the type of issue that each of the approaches could reveal. Observation-based techniques can expose the majority of issues relating to usability (CAT3), however the biometric-based approach enabled researchers to discover many more issues in categories related to players' feelings, immersion and gameplay experience (CAT1 and CAT2).

Methodology: This study considered six players with about 60 minutes of gameplay per player over two video game titles, which can be deemed a reasonable sample size from which valid conclusions can be drawn to provide formative feedback for game developers. Ideally, this work would be extended to include video game titles of differing genres, to further investigate the contribution of biometrics across game types. Whilst the study could be conducted with a greater number of participants, the post-session video analysis is highly labour-intensive and therefore a significantly larger sample would be impractical for the fast turnaround required in the games development cycle.

During the biometric-based approach, when prompted to recall elements of their gameplay experience, participants were infrequently not able to recall their thoughts or the circumstances shown to them. The experimenter was able to replay more of the gameplay footage if the player was not able to remember the particular moment. If the additional footage was not enough to facilitate recall the experimenter progressed to the following micro-event. Unrecalled events were excluded from the findings of the biometric- based approach.

The method proposed here (using GSR to structure post-session interview) is a novel approach to include physiological measures in GUR in order to conduct formative evaluation on gameplay experience. This is in comparison to previous studies (as discussed in section 2.4) where often focused on exploring correlations between collected physiological measures and self-report measures (such as questionnaires) as away of validating hypothesis in summative evaluation settings.

The Use of GSR: This study only used GSR, despite ease-of-access to further biometric measures. At the end of each gameplay session, the participant was asked about how they felt from wearing GSR sensor, and if the sensor had any impact on their ability to play the game, however none of the participants reported any difficulties in using the game controller. The low-intrusiveness of the electrode pellets when connected to the ring and little fingers allows the participant to quickly forget the presence of the sensors and does not severely impact the validity of the experiment. This is especially relevant when considering comparisons to EEG, which, for a reliable reading, would requires 16-32 electrodes attached to the scalp; and facial EMG, which requires electrodes to be adhered to the participants' faces. The sole use of GSR allowed us to minimise participant intrusion.

GSR provides an easy to collect and analysis source of data, with a fast response rate, reflecting the participants' arousal measurements in 2-5 seconds from the triggering video game event (Lang, 1995). Measurements of GSR are therefore highly suited to the live logging and also the live video capture procedures demonstrated in this study. GSR also provides a data format that is easily analysed, with clear indications of micro-events visualised from the raw data itself, and requiring no post-session review.

The use of finger-mounted sensors does introduce the problem of movement-induced signal artefacts. Throughout the study, a limited number of micro-events (as a result of arousal in biometric reading) were explained as signal noise, due to changing sitting position or stretching of the hands (both captured using the video camera in the study room), and therefore did not reflect players' biological responses to the game. These events were acknowledged during the event-logging process, and were not shown to the player during the explanatory phase. If a small movement went unnoticed during the logging process, either the video of the player captured during the gameplay session revealed the movement, or players reported that they could not recall the particular moment when prompted, and the experimenter proceeded to the following micro-event. The use of foot-mounted GSR sensors, or the application of an alternative biometric sensor that does not hinder the use of the hands, such as facial EMG or EEG, would reduce the number of movement-induced signal artefacts.

This study begins to highlight categories of player experience and usability issue types which evoke arousal in players' GSR levels, but there are issues discovered by the observation-based approach which remained undiscovered by GSR alone. Further research into the contribution of biometrics using differing biometric sensors, including those related to valence (such as EMG), may allow more of the player emotion spectrum to be represented, which may reveal latent usability issues.

Further Applications: The issues common to the observation-based approach and the biometric approach demonstrate the usefulness of biometrics as a validation tool. An equivocal usability or UX issue can be validated and confirmed by the presence of a biological response and player-reported confirmation of the problem.

Of the biometric findings, players explained many of the micro-events as positive gameplay experiences, but these positive findings have not been included in the analysis for this study. The focus of this study was solely on the negative issues, since negative usability or UX issues in video game titles are of particular interest to games developers for improvement purposes. Revealing positive events in video games under analysis may provide valuable feedback to game development companies, allowing them to quickly and accurately understand successful elements of their game. Physiological measurements have been used extensively to identify scenes of high arousal (e.g. (Drachen, Nacke, Yannakakis, & Pedersen, 2010a)); further studies in this thesis also discuss biometrics as a tool for the analysis, visualisation or validation of positive gameplay experiences. Indicating positive game events may therefore be considered to be as useful as finding negative issues and would contribute to the understanding of biometrics in GUR.

Although, this study did not focus on severity rating among identified issues, but, building on the results from this study, would provide basis for discussing the relative strengths of the different methods for video games user research.

The qualitative post-session interview results: The qualitative post-session interview results aimed to explore the helpfulness of using the proposed approach to understand players' behaviour. The interest of this study was not in mapping the results to inform precise design guidelines, but to show the effectiveness of this approach in understanding players' gameplay experience. As explained in the analysis section, this study was also not intended to interpret the arousal in GSR measurements to a specific emotional state, as well as a comparison between participants' biometric measurements to each other. Each participant's physiological data were only used to select micro-events from his own gameplay session.

The qualitative post-session interview results show that this approach provides another source to identify both positive and negative usability and user experience issues, which can be use in a triangulation setting with other methods to increase the confidence of user researchers in reporting issues. Indicating positive game events allows developers to better understand successful elements of their game, and can be considered to be as useful as finding negative issues.

The study by Gow, Cairns, Colton, Miller, & Baumgarten (2010) explores the value of using post-session interviews to provide a better understanding of player experience. In their study they asked their player to watch the whole recorded video of their playtesting session in order to describe their gameplay behaviour. If needed players could be prompted to talk by having access to a list of common experience words. Similar to other video analysis techniques, this is highly time consuming and would be less practical for a longer play test session. Researchers have implemented various approaches to facilitate participants' recall in post-session interviews, by using biometrics to point out significant moments in gameplay. The post-session interview can be made more efficient by only discussing those selected moments, which have been identified.

The qualitative post-session interview results focused on using biometrics to structure post-session interviews. The qualitative analysis presented here aims to point out the effectiveness and efficiency of this approach. In order to further evaluate this approach the quantitative result presented earlier detailed a comparison with traditional user testing approaches.

3.3 Study Two[4]: Physiological Arousal and Social Interaction Coding

After looking at how changes in player's arousal can be used as a manual book marking tool to identify events for post gameplay analysis, this section details a study which applies all the explained approaches (in section 3.1) to help GURs and designers to have a better understanding of player experience. The aim of this study is to demonstrate how these approaches can apply in a user test session to collect player data and cover different aspects of player experience.

The study presented here aims to provide an example of how analysing player's physiological arousal with behavioural coding enables game developers and user researchers to have a better understanding of player experience and their motivation in game. The impact of this study to the thesis is to enhance understanding of the contribution of physiological measures in GUR, and thus to reinforce triangulation of such approaches too as part of Biometric Storyboards.

This section starts by discussing the motivation for studying collocated video games social interaction, providing a brief overview of previous researches in the area, presenting the study's

[4] This study was presented at the multi.player conference (2011) and accepted for publication as a book chapter (see publication list P4). The study was designed and conducted by the thesis author in collaboration with S. Bromley (as part of his Master's studies (Bromley, 2011) at the University of Sussex). The thesis author conducted the analysis of physiological measurements and players' post-session interviews. S. Bromley developed the social interaction coding tool used in this study and conducted the analysis of social interaction data. The development of the tool used in this study is not part of the research presented at the thesis. This section is based on the book chapter.

settings and results, followed by a discussion of results and contribution for the field of GUR and in BioSt.

3.3.1 Motivation

Defining and measuring the forms of social interaction evident within collocated players is one of the current challenges for GUR, due to the result of the rise in popularity of multiplayer social and 'casual' games. The focus of this study is to present greater insight into player interaction during multiplayer collocated sessions, by developing previous social interaction research, combined with an understanding of player's intrinsic motivations by using physiological arousal and behavioural coding.

This study used the triangulation of physiological arousal measurements (GSR) with player post-session interview (the approach discussed in previous section), the social interaction coding tool (Bromley, 2012) and self-assessment diagram (as explained in section 3.1.1) to understand how forms of social interaction resonate with specific player types. These were then applied to a study of 16 players, across 8 sessions. The results of the study helped to develop an understanding of the motivations behind player's interactions during collocated gaming, which develops previous work on social interaction in multiplayer gaming. Moreover, it shows an example of how analysing players' physiological arousal in combination with behavioural coding can bring a greater understanding of their motivations to play games and how does this affect their experience.

3.3.2 Introduction

The continued success of casual collocated multi-player games, which rely on interaction between the players as a core design element is evident through the multiplayer focus of top selling games (e.g. *Wii Sports* or *Just Dance*[5]). As such, there is the potential to expand on the developer and researcher's ability to define and measure the forms of social interaction that occur during these sessions, in order to optimise these gameplay experiences.

Studying communication between players in collocated multi-player gameplay is key to understand player interaction and experience as the communication and interaction between collocated players often has impact on the overall gameplay experience. For example Drachen & Smith (2008) studied how format (media of expression) impacts on verbal communication in multi-player games; Voida, Carpendale, & Greenberg (2010) defined 6 key types of behaviour that emerged during social gaming sessions, and notes that their work paves the way for future research to examine the causal relationship between game dynamics and social interaction. This viewpoint is also seen in Bernhaupt's assessment of the current state of academic games

[5]http://www.gamasutra.com/view/news/35893/Ubisofts_Just_Dance_2_Becomes_Wiis_BestSelling_ThirdParty_Game.php

research, where it was noted that steps needed to be taken to bridge the divide between academic research and commercial games development (Bernhaupt, 2010). Hence this study aimed to develop research in defining the types of social interaction in collocated gamers, and identify which behaviours cause positive reactions in gamers.

This study therefore has two contributions; primarily, it demonstrates an iteration of pre-existing work on defining social interaction, tailored to a specific genre of interest to commercial games development. The second aim (specifically for this thesis) was to better understand the contribution of physiological measures in GUR. Where it demonstrates how various user research methodologies can be used in a triangulation setting to provide the researchers and game developers with an enhanced understanding of various aspects of player experience.

The study presented here concurrently records biometric responses (section 3.1.2) alongside a chronological record of the social interaction present (Bromley, 2011), and applies this to a social gaming setting. As discussed earlier in GUR methods, the insight gained from qualitative user testing is often limited by the effectiveness of for example post-session interviews, due to the delay between the gameplay experience and the player's recollection of it. The methodology presented in this study was designed to triangulate multiple sources of data to tackle this situation.

3.3.3 Social Interaction

Voida et al. (2010) made advances in defining 6 key types of behaviour displayed in collocated gaming sessions, through a review of player's behaviour during multi-player games such as Guitar Hero and Mario Party. In their research they ran a number of collocated group console gaming sessions, and identified behaviours which related to, or altered the social dynamics of the group. Through questionnaires and group gameplay sessions, they identified the key forms of interaction, which were utilised for this study, as the basis for coding social interaction.

Determined from an iterative test and evaluation process performed with Relentless Software, and the interviews with the development team (Bromley, 2011), this study utilised revised categories based on the development priorities as described in Table 3-3 (next page).

Voida et al.'s Category	**Revised Social Interaction Category utilised in this study**	**Description**
Constructing Shared Awareness	Shared Awareness	Shared Awareness includes building a shared awareness of the game state, and can include collaborative working out, giving hints, or making another player aware of something within the game, such as game mechanics or "what to do". It can also include reporting to other players what activities you are performing within the game.
	Requesting Information	Requesting Information typically includes asking about what is happening in game, how the game works, or how to achieve their goal. It can also include asking other players to report their status. It is often combined with a period of shared awareness.
Reinforcing Shared History	Shared History	Shared History includes discussing what happened earlier in the game, or in a prior play session. May include links to other games, or with players not present.
Sharing in Success and Failure	Shared Success	Shared Success includes celebrating a group success, or congratulating another player on their success. It can include a group celebration despite being in a competitive situation.
	Shared Failure	Shared Failure includes taking group responsibility for failing a task, offering reassurance, or commiserating with a player who has failed a task. It does not include blame (which may be more appropriate under trash talk).
Engaging in Interdependence and Self-Sacrifice	Team Optimisation	Team Optimisation includes discussing the group dynamics, or negotiating an individual's contribution to the group. It can include assessing the ability of others, and discussions over who is leading or in control. It can also include denying players the chance to join in.
Talking Trash	Trash Talk	Trash Talk includes celebrating your own success over the other players, or laughing at their failure. This can be in competitive or collaborative game types, and often involves put downs or insults.
Falling Prey to the computer's holding power	Self Indulgence	Self Indulgence includes not playing the game at the expense of other player's enjoyment, making up their own meta-game or not participating fully, leading to a disruption of the flow of the game. It can include repeatedly performing the same action (i.e. viewing a hidden in-game feature, or 'Easter egg').
-	Off topic	To capture any interaction that was non-game related.

Table 3-3 Showing Voida & Greenberg (2009) categories of social interaction behaviour, and the adapted categories (Bromley, 2011) used in this study.

3.3.4 Player Profiling

Previous researches in player satisfaction models, such as (Nacke, Bateman, & Mandryk, 2011), demonstrate that player's reactions to events were not uniform, with some players reacting strongly to 'winning', whereas others had strong reactions to social interaction. Hence, it was noted that an approach required for classifying players to generalise their reactions and draw specific conclusions would be applicable to professional developers.

For this reason, this study utilised online Bartle's Test of Gamer Psychology (Andreasen & Downey, 2003), created based on (Bartle, 1996) research. The test presents 30 dichotomous questions centred on each player's intrinsic motivation in online games, and aims to categorise players of online role-playing games (RPGs) into four groups, as below:

- Killers (Clubs): are interested in combat/competition with other human players, and prefer this over interaction with non-player characters.
- Achiever (Diamonds): are most interested in gaining points or alternative in-game measurements of success. These players will often go out of their way to gain items that have no in-game benefit besides prestige, such as 'achievements' or 'trophies'.
- Explorer (Spades): These players are interested in discovering the breadth of a game, and will explore new areas or take non-optimal routes to explore. They do not like time limits, since this limits the potential to explore options.
- Socialisers (Hearts): These players are interested in the social aspect of gameplay, rather than the game itself. They enjoy interacting with other players, and use the game primarily as a means of communication.

Through analysing players based on their player type, this study could capture and represent player's motivations whilst playing games, and how this affects the types of social interaction evident during gameplay.

3.3.5 The Study

The Games: Buzz! Quiz World (Relentless Software, 2009), was chosen as the game for the study as it was deemed a suitable game to encourage social interaction. The game takes the form of a multiplayer trivia quiz, lasting up to 30 minutes, where each player is given multiple choice questions to answer, using a custom controller that allows the quick selection of the player's choice (Figure 3-8). Earlier rounds in the game allow the players to score points, which give the player an advantage in the final round. One player is declared to be the winner.

Figure 3-8 The Buzz Controller

By putting the players into direct competition, while allowing them to compete simultaneously, a pilot study implied that 'Buzz! Quiz World' would create a diverse range of interaction, and would allow a qualitative evaluation of the differences between different player types. As a trivia game, it also relied little on manual dexterity or knowledge of an intricate control system, and hence would minimise any bias as a result of player's previous experience with the game, reducing variables in the study. Gameplay could also be performed with only one hand, ensuring that the GSR signal suffered minimal disruption from the player's hand movement.

Participant selection was based on a number of criteria. In order to ensure that the study recreated authentic social interaction, as would be found in a typical collocated gaming session, pairs of players from pre-existing social groups were recruited. This would ensure that the player's reactions would be equivalent to a social interaction experience found in normal gameplay, and reduce any bias or -typical behaviour introduced by playing with unfamiliar opponents.

The use of pre-existing social groups for studying group dynamics has been previously established by Mandryk, Atkins, & Inkpen (2006), in their work on modelling player's emotional state based on physiological response. Through the comparison of games involving both collocated friends and collocated strangers, they noted that there is a variation in player's physiological reactions to in-game events based on who they were playing with.

Prior to each session, each player was asked to perform an independent self-evaluation, based on the descriptions of Bartle's player types, to self-classify their player category. This was then compared to their result from Andreasen and Downey's online adaption (Andreasen & Downey, 2003) of Bartle's quiz, to validate their self-selection, with the results of both recorded. As noted in the discussion section, all players agreed with their assigned category. This was performed for a number of reasons; it ensured that recruitment covered all four player types (but not in all possible combinations) and hence represented a wide player-base. It also enabled the

results to be broken down by player type, and hence greater insight into the characteristics of different player types could be gained in the post-session analysis. Despite the quiz focuses on Online RPG's such as Multi-User Dungeons, the distinctions seem appropriate for a wide range of multiplayer games, as social interaction is also evident in collocated social games. As such, this study bases its categorisation of players based on these criteria, particularly due to the ease of administering the test based on Andreasen and Downey's online adaption.

This was performed independently of the researchers, who were unaware of each player's assigned type during the sessions and while categorising the social interaction evident between players, to prevent this biasing the analysis of player's interaction.

Study Design: The study introduced physiological measurement and post-session interviews (as explained earlier) in order to gain greater qualitative data on the motivations behind player interaction. By using this to guide interviews with players after the gameplay session, it became possible to understand not only what in-game events players were reacting to, but why they were responding to them.

During the pilot study, each player was asked to play a short single-player mode individually prior to the start of the multiplayer session. The intended goal of this was to ensure that bias was not introduced due to a player's unfamiliarity with the nature of the game, or with the control method, and that each player had an equal chance of succeeding within the game. However this was not repeated in the full study. It was noted that the primary usability issue encountered by players was the lack of understanding of when it was each player's turn to answer the question, and this idiosyncrasy only appeared in the multiplayer mode; Hence a single player pre-session did not reduce the occurrence of this usability issue. This was replaced by a briefing prior to the session beginning, explaining the rules of the game, how the controls work, and when players should answer. Since the focus of this study did not rely on a 'fair game', this was judged an appropriate method of overcoming potential usability barriers. This also gave the advantage of capturing the 'new player' experience, and hence gave another area for analysis in the final study, as has been explored in the results section.

The players were required to play one game of 'Buzz! Quiz World', which comprised of 6 rounds, lasting up to 30 minutes. The rounds were all variations on trivia questions, requiring the player to answer the question correctly; however each round had individual idiosyncrasies in how the player was scored. The rounds have been captured in the participant table, and are as follows:

- Point Builder (Opening Round): The players were given points for correct answers, with no time limit.

- Fastest Finger: Players are given points for correct answers, with the player who answers first gaining a larger point bonus.
- Stop the clock: Similar to Fastest Finger, the players are given points for the correct answers, based on the speed of their response.
- Boiling Point: Each player gets the opportunity to answer questions, with the first reaching 6 correct answers winning a point bonus.
- Pie Fight: Upon answering a question correctly, the player must choose a player to throw a pie at. Correct timing is required to avoid the player throwing the pie at themselves, and hence losing points.
- Over the Edge: Answering a question incorrectly elevates the player over a pool of gunge. Each incorrect answer brings them closer to the gunge. After 5 incorrect answers, they are thrown into the gunge, losing that player points, and ending that round.
- Pass the Bomb: One player is given a bomb, on a timer of unknown length. They are then asked a question. Correctly answering a question passes the bomb to the second player, who is then also asked a question. After an indeterminate time, the bomb will explode, losing that player points.
- Final Countdown: Each player is elevated based on their current point score. They then slowly descend towards the ground, while being asked questions. A correct answer will raise your platform up slightly. When one player touches the ground, they are out.

Each game comprises of the opening round, the final countdown, and a random selection of the other rounds. Social interaction data was also noted in between rounds, during the review of the results from the previous round, and the selection of category for the following round. Where appropriate, player type's reactions to individual rounds have been noted in the results section. Table 3-4 shows the sessions, players, and rounds played.

The study was conducted with the similar setting as the study S1 and facilitated the use of GSR, as measurements of GSR are suited to the live logging and also the live video capture procedures demonstrated in this study. GSR also provides a data format that is easily analysed, with clear indications of micro-events visualised from the raw data itself, and requiring no post-session review. The use of finger-mounted sensors does introduce the problem of movement-induced signal artefacts; however this is mitigated by the custom Buzz controller (Figure 3-8), which requires little movement from the player to utilise, and could be operated entirely with one hand.

Session	Player A (Category)	Player B (Category)	Rounds Played
1	A1 (Achiever)	B1 (Killer)	Point Builder, Beat the clock, Over the Edge, Short Fuse, Fastest Finger, Final Countdown
2	A2 (Killer)	B2 (Killer)	Point Builder, Pie Fight, Boiling Point, Fastest Finger, Short Fuse, Final Countdown
3	A3 (Explorer)	B3 (Explorer)	Point Builder, Stop the Clock, Boiling Point, Fastest Finger, Short Fuse, Final Countdown
4	A4 (Achiever)	B4 (Explorer)	Point Builder, Pie Fight, Over the Edge, Short Fuse, Fastest Finger, Final Countdown
5	A5 (Achiever)	B5 (Socialiser)	Point Builder, Boiling Point, Stop the Clock, Fastest Finger, Short Fuse, Final Countdown
6	A6 (Killer)	B6 (Achiever)	Point Builder, Over the Edge, Boiling Point, Short Fuse, Fastest Finger, Final Countdown
7	A7 (Killer)	B7 (Explorer)	Point Builder, Boiling Point, Stop the Clock, Fastest Finger, Short Fuse, Final Countdown
8	A8 (Killer)	B8 (Socialiser)	Point Builder, Boiling Point, Pie Fight, Short Fuse, Fastest Finger, Final Countdown

Table 3-4 Session data showing the players and rounds present in each game

Throughout the study, a limited number of micro-events (as a result of arousal in biometric reading) were explained as signal noise, due to changing sitting position or stretching of the hands, and therefore did not reflect player's biological responses to the game. If a small movement went unnoticed during the logging process, either the video of the player captured during the gameplay session revealed the movement, or players reported that they could not recall the particular moment when prompted, and the experimenter proceeded to the following micro-event.

During user test sessions, the players were remotely observed and the researcher noted any verbal interaction between the players and classified it within the revised categories using the coding too based on the categories described in Table 3-3. The tool automatically (Figure 3-9) time-stamped the start and the end of each noted interaction. As a result of performing the

behavioural sequential analysis, the coupling between the player's behaviour and resulting experience (if any) is easily visible as a temporal pattern.

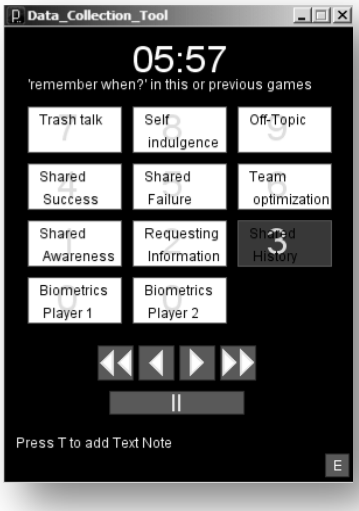


Figure 3-9 The social interaction recording tool interface, displayed recording a Shared History behaviour

After the session, participants were asked to perform an unguided review of the 'player experience' of the session (self-assessment diagram, Figure 3-1). To evaluate this, they were given graph paper with the x-axis indicated as time, and the y-axis as 'player experience'. Participants were then asked to plot their experience over time, and their own recollection of their feelings. This task was devised to gain some insight into their own impressions of the session, prior to receiving any mental 'cues' through the rest of the evaluation process, and was intended to give some insight into what parts of the game particularly resonated with players after the session.

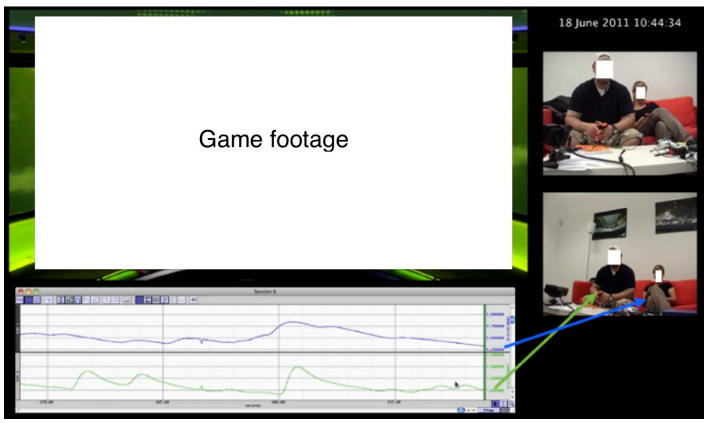


Figure 3-10 Observation screen for S2

To gain insight into how they felt during the session, the players were then required to view the video data from the full 30 minute session with a researcher. This video data included the game screen, a visualisation of the GSR data and the player's own physical reactions (Figure 3-10). The players were asked to give feedback or their interpretation of what provoked peaks in the GSR signal, or when occurrences of verbal interaction were noted. This was guided by the researcher who would prompt for the player's insight into their feelings when any anomalous behaviour or data was noted.

There were a number of reasons for choosing this methodology as an appropriate way to gain an insight into the player's experiences whilst playing the game. Players were required to watch the entire session, not just the moments with biometric peaks, to give them the context in which these events happened (this was a different approach to study S1 as the nature of a puzzle game makes it difficult for players to remember the sequence), and allow them to verbally feedback a truer understanding of what they were thinking at that time.

This methodology was also favoured over a think-aloud study, where the players would give verbal feedback during the session to improve the quality of the data. Unnecessary verbal interaction creates an artificial mental overhead during the play session, which would not truly represent a collocated gaming session. It was also noted that during multiplayer sessions players often talk about the game without requiring a formal methodology and this would prove a closer match to true social interaction.

After the sessions, the social interaction data was analysed quantitatively, to gain an understanding of the characteristics of each player type. This was then triangulated with qualitative data, gained through the user's self-evaluation, and the biometric-based post session interviews, to give a depth of understanding to the data, and explain why players acted the way they did.

3.3.6 Results

It was noted that the player's self-expressed motivation, biometric 'peaks' and types of social interaction closely aligned, and implied that players within each category shared characteristics during social gaming sessions.

Table 3-5 shows the outcome of the coding tool and the extent to which each form of interaction was noted during the gameplay, as a percentage of each session's total interaction.

			Percentage of total interaction in the session								
Session	**Player A**	**Player B**	**Off Topic**	**Shared Awareness**	**Requesting Information**	**Shared History**	**Shared Success**	**Shared Failure**	**Trash Talk**	**Team Optimisation**	**Self Indulgence**
1	Achiever	Killer	2	13	9	28	8	13	23	0	4
2	Killer	Killer	2	4	16	24	9	18	27	0	0
3	Explorer	Explorer	1	7	20	34	10	15	13	0	0
4	Achiever	Explorer	0	26	19	24	5	15	11	0	0
5	Achiever	Socialiser	0	28	10	15	26	15	6	0	0
6	Killer	Achiever	2	45	15	15	0	23	0	0	0
7	Killer	Explorer	0	60	3	5	9	18	5	0	0
8	Killer	Socialiser	0	42	10	18	5	15	10	0	0

Table 3-5 Table showing the percentage of the session's total interactions in which each type of social interaction was noted.

By triangulating these quantitative studies with the qualitative data gained through the player's self-assessment of their experience and the structured post-session interview noted by the GSR arousal, it is possible to see patterns in the biometric and interaction data, when the players are divided into their player 'type' categories.

It's also worth noting that these sessions saw no instances of team optimisation, and very little of self-indulgence. Due to the sessions involving only two players, in direct competition, there was little potential for team optimisation (such as taking other player's turns for them) to occur. Self indulgence was also likely to be minimised, due to the novelty of the title, short length of the sessions, and the formal setting (rather than playing at home!).

Killers: As evident in Table 3-5, sessions which involved the killers showed a much higher degree of 'trash talk' interaction than games without killers. This trend is particularly evident in the session (number 2) with two killers playing one-another, which was the game with the highest degree of trash-talk. The exception was session 6, where the killer player A6 didn't display any trash talk during the session at all. As elaborated upon in the discussion section, this is likely to be due to their pre-existing social dynamic, since these players were married. As

discussed later, the biometric results still implied that player A6's intrinsic motivations still fit within the killer archetype.

Sessions involving the killers also had the highest occurrence of 'shared failure', the behaviour which indicated discussing and taking shared responsibility for failure in a round, or getting questions wrong.

This data was supplemented by an understanding of the key in-game actions and interactions that lead to strong GSR peaks. It was noted that the killers' arousal was highest when the game was competitive, and against opponents perceived of being of equal skill. When they became aware that they were going to win, or that the opponent showed little competition, their levels of GSR arousal dropped rapidly.

Arousal for the killers was also heightened at the start of the new rounds, and when they were asked questions. Potential explanations for this will be discussed in the discussion section.

Achievers: The quantitative analysis of the data showed that the 'shared history' interaction type was most prevalent among groups that included achievers. The shared history behaviour is the act of describing or reliving previous game experiences, such as how the previous round went.

It was also apparent through the achievers' interaction that they were more likely to share their answers in the sessions, when they believed they had the correct answer. Despite the advantage this potentially gave to the other players, the achievers continued to demonstrate that they knew the correct answers throughout the session.

Depth to this interaction data was provided by the analysis of their GSR responses. This revealed that the achievers showed a higher degree of arousal, both through their GSR responses and their self-assessment, in the 'Over the Edge', 'Short Fuse' and 'Final Countdown' rounds. Arousal level for the achievers was also highest during the explanation of rounds and in-game instructions. The reasons for this are discussed in the next section.

Socialisers: The quantitative interaction analysis showed that the sessions which involved socialisers displayed the highest amount of shared awareness, where the players would discuss what was happening within the game. Socialisers were also the primary cause of the 'shared success' behaviour type, where successfully answering questions or winning rounds, resulted in congratulations and joint celebration, even if the player hadn't personally won.

It was also clear through the socialisers' interactions that, similar to the achievers, it was common for them to share answers during the sessions. However it was noted that the nature of

this interaction was slightly different, with a focus on collaboratively working out an answer, rather than displaying the player's own knowledge.

The use of biometrics and user interviews revealed that during these sessions the socialisers did not display strong intrinsic reactions to the progress of the game, whether successful or not. This is distinct from the other player types viewed, where success in a round would typically show a significant biometric reading. The ramifications of these observed results are discussed later.

Explorers: Sessions that featured explorers displayed the greatest degree of requesting information, where one player would be asking about a game mechanic, or what was happening, than games without explorers. They also displayed the highest degree of the shared history behaviour type, with this behaviour being equally prevalent in games featuring explorers and achievers.

During these sessions, explorers showed a low degree of competitive interaction, such as trash-talk, with one explorer (player B4) giving the opinion that "*it didn't matter if I win or lose*". Biometric data gave an insight into the in-game events which created a strong emotional response. Throughout the sessions, the explorers displayed relatively static reactions to events, with slight peaks when being asked challenging questions or exploring new topics.

Other in-game events, which lead to significant biometric responses include an increased arousal when players were selecting the round themselves, or when being asked new questions in an unfamiliar topic. Unlike players from the other groups, they also displayed an interest in the in-game animations of the player's avatars, and the aesthetics of the game environment. The discussion section gives some depth on the observed behaviour noted, and offers potential explanations for why this is the case.

3.3.7 Discussion – What Does This Mean for Game Developers?

The results show correlation between the interactions and biometric results emerging among players who share categories. By combining this with an understanding of player motivation gained from user interviews, player profiles have been developed for each 'type' of player.

For killers, it was seen that GSR arousal was highest during these sessions when the competition was perceived as '*worthy'*, and the players were close in skill level or points. Arousal sharply dropped when it became obvious that the killer was going to win, and this is very closely reflected in the player's self-assessment graphs, such as one in Figure 3-11. Hence the primary recommendation for targeting games development towards killers would be to use mechanics to ensure that both players always have an ability to win – as evident in the Final Countdown (final round of 'Buzz! Quiz World'), where the points from the previous rounds give one player an advantage, but do not prevent the other player from winning.

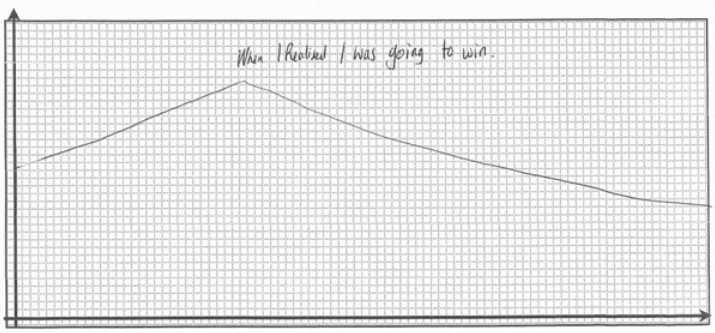


Figure 3-11 A Killer's self-assessment graph (player B2)

Throughout the sessions, it was also noted that trash-talk and competitive interactions were the primary form of interaction used by killers, and this idiosyncrasy could be developed by in-game mechanics, such as the ability to taunt, as found in other multiplayer games.

Among the achievers studied, increased arousal was noted in rounds where there was a visible indication of progress. For example, the 'Over the Edge' round raised players based on their progress, and the loser's avatar was visually punished at the end of the round. Through self-assessment, GSR peaks, and post-session interviews, it was evident that this round, as well as the other 'visual' rounds such as 'Short Fuse' or 'Pie Fight', resonated particularly with achievers.

Achievers requested additional visual displays of progress or ability levels, as evident through player A5's comments suggesting that the audience should be shown to support the players by "*shouting their name when players are doing well*". Therefore an additional recommendation for development to target achievers would be to increase the degree to which progress or success can be displayed during the game, and allow players to display their skill through badges or trophies.

It was also interesting to note how achievers were prepared to sacrifice the primary goal of the game, winning the quiz, in order to demonstrate their knowledge of the answers. Achievers, such as player A1 and player A4 were frequently noted to be answering questions out loud, despite the advantage this gave their opponent. It can be concluded that displaying the breadth of their knowledge is a more important priority to achievers than winning the game.

In a comparison between achievers and killers, it was noted that killers showed increased arousal when asked questions, whereas achievers showed biometric peaks when instructions were being given in game. A potential explanation of this would be that these player's intrinsic goals have slight differences – killers in these sessions were interested in getting individual questions right, and the opportunity to defeat the other player, whereas the achievers were interested in learning the game mechanics in order to understand how they could meet the in-

game criteria for visible success, evident through the changes in when GSR arousal was noted in the timeline of the gameplay sessions.

Socialisers were noted to have a low level of arousal towards the progress of the game, and did not show GSR increase as a result of in-game success. Instead it was obvious through their interaction data that they were most interested in shared experiences, and hence showed prominent 'shared success' interactions, where they would congratulate their opponent.

As noted in the results section, shared awareness was also evident with socialisers; however, unlike achievers, the nature of this interaction seemed more collaborative than 'showing off'. As such, it can be recommended that games should give player's opportunities to discuss in-game events – this was particularly noted during 'Buzz! Quiz World' at the end of rounds, where the results of player's in-game positions were talked over by socialisers, since the data was already known. Additional 'down-time' to discuss the game would aid in targeting players who display socialiser characteristics.

In contrast to other player types, interaction among explorers was largely noted to be co-operative and inquisitive, such as describing the animations displayed by the in-game avatars, "*look he's hanging himself*" [B3]. They would work with their opponent to select the rounds that would be of most interest to both, and showed GSR arousal peaks when the questions were of interest. Their verbal interaction was largely discussing the question's topics or the game's mechanics. As such, the recommendation can be made to target games towards explorers through offering a non-repetitive gaming experience, with a range of forms of interactions and avenues for exploration, such as a variety in the rounds and topics. This motivation was also noted in the comment from an explorer that the rounds Point Builder, Fastest Finger and Stop the Clock were too similar.

In the majority of players, the results in the online test matched their self-assessed 'category', and when players were asked whether they felt their assigned group represented how they played games, most players indicated yes. Some players felt that they fell into two categories equally, but no player outright disagreed with their assigned category.

This study also indicates that player's behaviour falls into categories – the biometric data has shown what aspects of the game, and what forms of interaction, resonate with players. As such it is possible to create a deeper understanding of the intrinsic motivations of each player type, with the biometric data giving greater insight into what aspects of gaming shows particular resonance with killers, achievers, socialisers or explorers. This study has also given an understanding of what types of social interaction resonate with each player type, and hence gives an increased ability to target in-game events to specific demographics.

This study develops the previously identified behaviours by giving greater understanding of the associated emotional context to each behaviour, and establishing a causal relationship between specific game dynamics and player engagement.

This study has also given greater depth to the aspects of social behaviour previously defined. By linking particular behaviours to specific types of players, an increased understanding of the nature of these interactions can be achieved. A causal relationship has been identified between specific game mechanics and occurrences of social interaction among player types, such as the occurrences of trash-talk among killers as they defeat an opponent close in skill level, or the use of shared awareness by achievers to promote their own success. This adds to the existing knowledge of social behaviour types by defining where each form of interaction is likely to be observed, and by whom.

It was also interesting to note that, based on pre-existing social relationships, players could appear to act typically to their 'type'. For example an in-game 'killer' was not overtly competitive whilst playing against his wife. However in these cases, their biometric results gave a clear indication as to what aspects of the game truly engaged them, as was later verified through user interviews. The use of GSR data to indicate player's interest towards in-game events hence allowed a much greater degree of insight than would have been possible through the sole use of measuring social interaction, or post session interview techniques.

The study highlighted how interaction data emerged between groups of friends playing social games together. It would also be of interest to see if these patterns were replicated when playing with strangers, or whether the interaction data would be emphasised, and players would display a greater amount of behaviour within their archetype. There would also be the potential to explore a wider range of biometric data, such as using facial EMG sensors to gain a greater understanding of player's reactions to social interaction behaviour.

The application of this methodology to other genres may also discover further applicable lessons for games development. The trivia game 'Buzz! Quiz World' was chosen to reduce the potential of variables surrounding skill or practice influencing the study; however another potentially interesting avenue to explore would be how interaction works in skill and reflex based games. In particular, it would be interesting to see if the division between player types remains as clear in a directly competitive game, or whether an increase in competitive interactions, such as trash-talk, would be noted across all player types.

By ranking the extent to which each form of interaction caused a significant GSR response, it has been possible to identify the degree to which each aspect of social interaction resonates within different psychographic groups, and hence which behaviours developers should be

encouraging in order to design exciting and engaging experiences that create strong emotional resonance.

This study has provided a deeper insight into the understanding of how players interact while playing games, and has shown that behaviour patterns emerge when the players are categorised by their implicit goals. However there are several potential extensions to this study which would help give a greater understanding of player behaviour in collocated gaming. This study solely looked at players in groups of two, and the interaction between two different player types. It would be of interest to compare these results to games involving larger groups of players. Larger group dynamics may give the potential for increased social interaction, and it would be of interest to see if the nature of the interaction changed too. In addition, Drachen & Smith (2008) described how interactions from each player type differ based on the player type of their opponent – there is much potential for further work defining the shift in social interaction evident across a range of opponents.

This research has presented methodologies for recording biometric responses and interaction data in order to gain greater insight into the player experience. It's been possible to see idiosyncratic GSR reactions to in-game events and the display of different forms of verbal interaction and players' post-session interview comments. This motivated the development of BioSt as a tool to visualise these relationships.

3.4 Summary

This chapter discussed the approaches that were developed from utilising phasic physiological measurements in conjunction with post-session interview, and detailed two studies where these approaches were applied in conjunction to other user research methodologies to enhance the understanding of player experience.

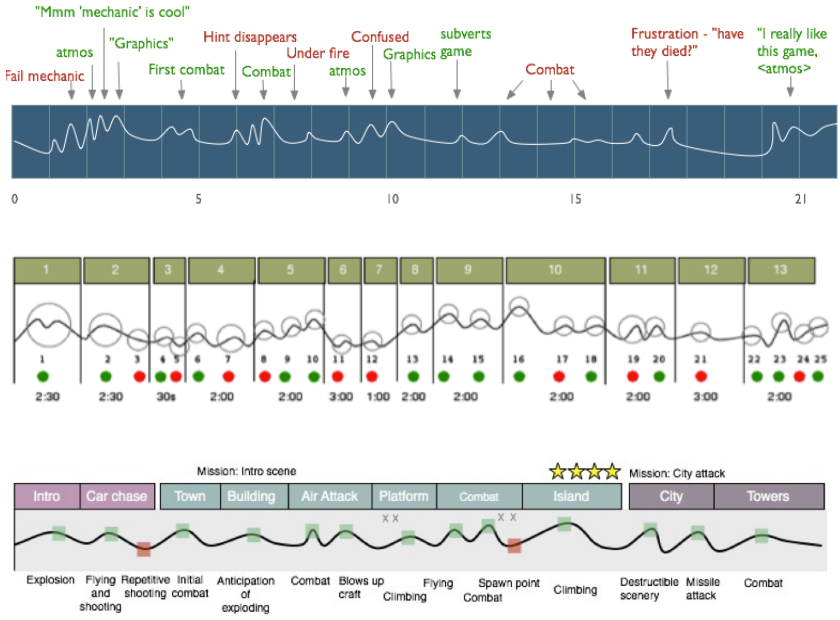
Through these two initial studies the contribution, advantages and limitations of physiological measure in GUR has been explored. As a result the coupling between the player's behaviour and resulting experience (if any) is noticeable as a temporal pattern. However, due to the limited visible area of a screen, it's difficult to get a visual overview of the complete player experience. The idea of Biometric Storyboards was then developed to overcome this issue and allows the entire player experience to be scanned quickly. In addition to this, the result of studies presented in this chapter iterated the data gathering methods later used for BioSt.

Chapter 4 goes on to describe Biometric Storyboards, explaining the creation and evaluation of three prototypes, for working towards the basis for the BioSt method and tool.

4 Biometric Storyboards: The Prototypes

This chapter explains the development of the Biometric Storyboards technique to visualise the player's experience journey over a longer gameplay period. This chapter details three case studies with under-development commercial games subjected to a series of user test sessions. The results of these user test sessions were presented to the game publishers using BioSt as well as traditional text and video reports. Each case study resulted in a prototype of BioSt and the feedback from the game publisher taken into account to iterate each prototype. Hence, three prototypes (F1igure 4-1) of BioSt have been developed and evaluated based on these case studies.

This chapter also details a further evaluation of the three prototypes through semi-structure interviews with game development professionals. The results from these evaluations established the requirements for further development of the BioSt tool.



F1igure 4-1 Biometric Storyboards; top: first prototype, middle: second prototype, bottom: third prototype

4.1 An Iterative Design Cycle

Evaluating and communicating user experience in games is an important component of the growing field of GUR. However, if using a physiological based evaluation, a major challenge for game industry and researchers alike is tying physiological measures and player experience

reports together, to provide insights because of the different data quality. BioSt allows GUR professionals to visualise meaningful relationships between a player's physiological changes, the player's self-reported experience, and in-game events. This chapter explains how BioSt was developed iteratively, by running three case studies with game design studios and by interviewing game developers about BioSt's advantages and disadvantages. Refining these prototypes led to the development of the BioSt tool (explained in Chapter 5), a software application that aims to facilitate reporting and analysing user experience issues for video games under-development.

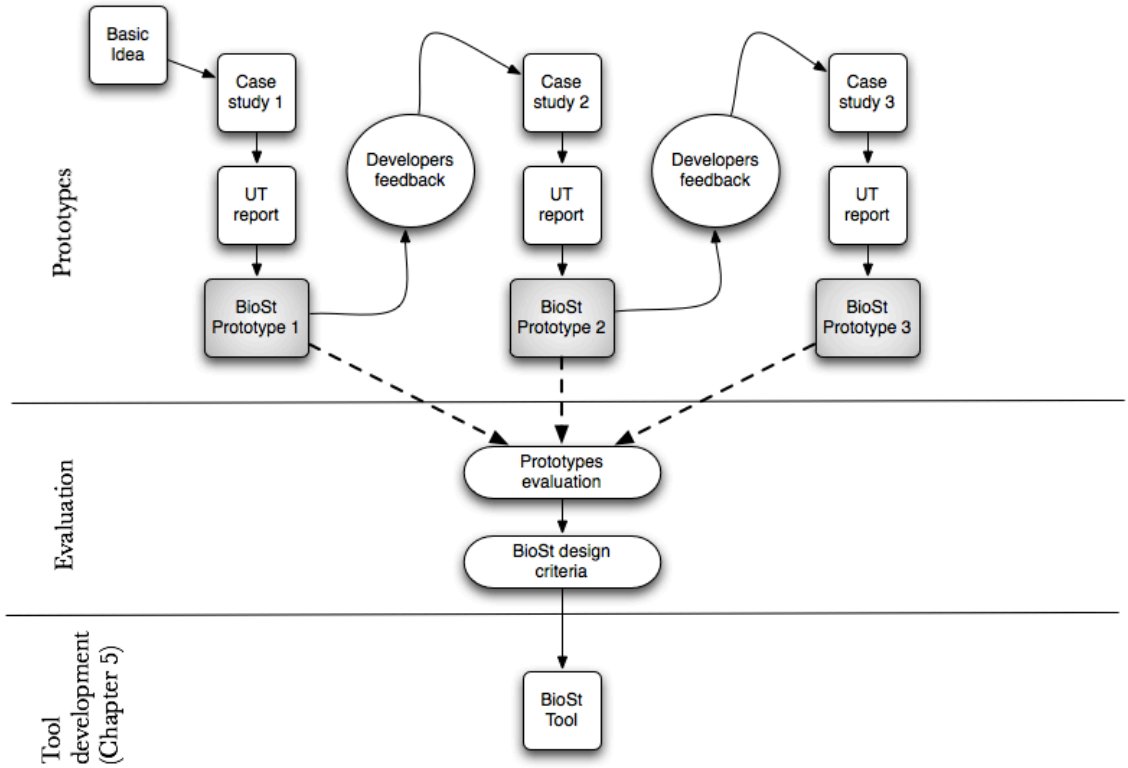


Figure 4-2 BioSt design cycle

As shown in Figure 4-2 the design cycle of the BioSt visualisation went through four phases over three years:

- Refine the initial idea in iterations while conducting three case studies in collaboration with game development studios. (Sections 4.3, 4.4 and 4.5)
- Evaluate the early BioSt prototypes by presenting them to game developers and interviewing them about advantages and disadvantages of this technique. (Section 4.6)
- Developing the final tool. (Chapter 5)
- Evaluation of a game designed using BioSt tool together with game programmers and independent game developers. (Chapter 6)

4.2 Introduction

Player experience is difficult to evaluate and report, especially using quantitative methodologies in addition to observations and interviews as experience factors (such as fun, enjoyable) are hard to measure and quantify. Most classical user research evaluation techniques (such as surveys and questionnaires) do not simply map to player experience, due to the engaging and fluid nature of games. As discussed in Chapter 2 and 3, one of the challenges in the quantitative player evaluation is to be able to collect data from users (in our case players) without interrupting their gameplay (continuous and unconscious). Since games also thrive on emotional experiences, the physiological evaluation is becoming an accepted method together with traditional interviews for player evaluation. However, one of the challenges is making the interpretation of physiological and player evaluation data meaningful in terms of facilitating design decisions for developers.

Data gathered in physiological evaluation studies is hard to explain and communicate to a games design team. This thesis is a step towards tying quantitative physiological measures of players to qualitative data from player experience reports. Steps in this direction are necessary to facilitate the interpretation of these large datasets, possibly creating visual aids for faster navigation and easier interpretation of physiological data. To meet this need this thesis focuses on the development of a player evaluation approach called Biometric Storyboards, where the player's gameplay experience is graphed over a longer period (e.g. a level of a game or several hours of play) based on using a player's physiological measurements and stories (or storyboards technique).

4.2.1 Storytelling

There are many purposes for employing storyboards to convey a narrative behind several aspects of game development, such as gameplay, art, animation, marketing, and the actual game narrative. Gingras (2012) talked about the use of storyboarding to find a shared vision for developers in all these areas. For example, stories often follow a pattern of high and low tensional events and dramatic cues to tap deep into an audience's collective psyche.

Narratives have always been part of the user experience process to communicate how and why a design would work (Quesenbery & Brooks, 2010). To leverage the power of narrative, the BioSt approach is based on using storytelling (or storyboards). There are many purposes for employing storyboards to convey a narrative behind several aspects of game development, as there is a pattern behind great stories. To be entertaining for example, stories need to have the right dramatic cues and tap deep into an audience's collective psyche. (See Figure 4-3 for an example sketch of a story arc)

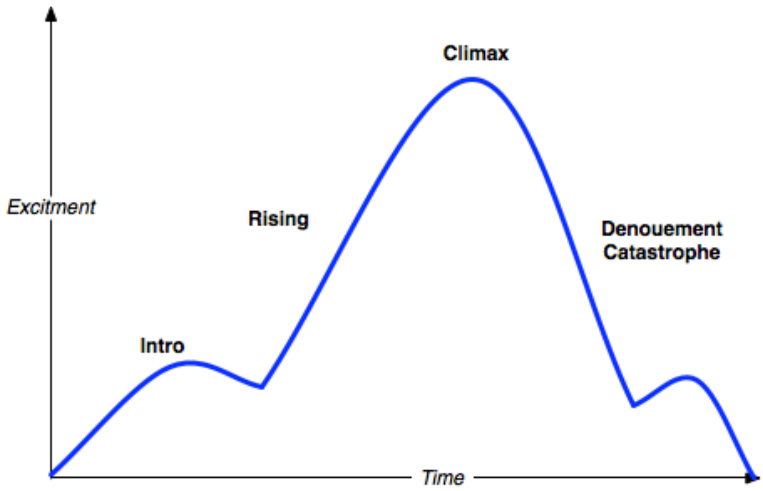


Figure 4-3 Example of story arc picture

Physiological measures, in particular GSR, can provide a suitable assessment of a players' level of excitement or frustration (arousal) and these measures seems to be useful for motivating the change in the players excitement graph (Figure 4-4).

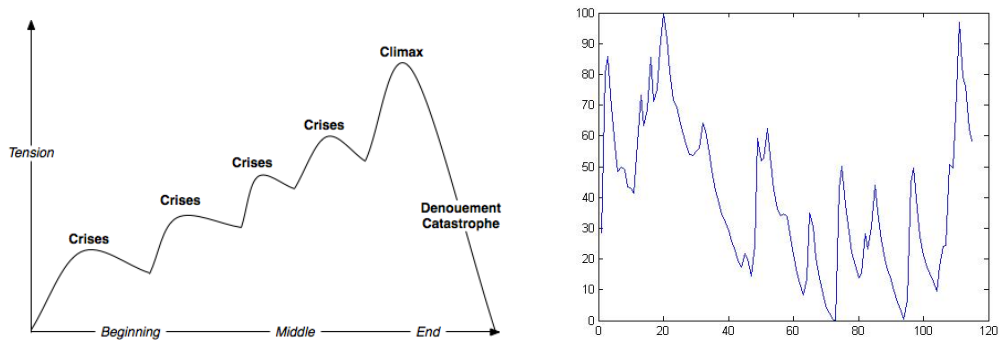


Figure 4-4 Example of a story arc (left) compared to an example of a player's GSR over time (right).

On the other side, storytelling is at the heart of the human experience. Once we experience something great we often cannot wait to tell someone about it. Stories help us shape our perceptions and consolidate our feelings. User experience researchers have leveraged the power of storytelling to drive observation-based and focus group research for improving websites and interface designs such as (Crothers, 2011; Inchauste, 2010).

This section argues how and why games development (or more specifically GUR) can benefit from storytelling techniques. For example in scenario based design (Rosson & Carroll, 2009) textual narrative descriptions of an imaginary situation are employed in a variety of ways to guide the development of an interactive system; in video games development the data gathered during user test sessions can make more persuasive stories (or scenarios). In BioSt the focus is

on stories that have the goal of describing and communicating player experience aspects to the game development team.

Storytelling is one of the most natural and powerful ways to share information. As part of the user experience, stories help designers to put the work in a real context and show design concepts or connect ideas. But more importantly stories help to keep users in the centre of the design process. This is critical when developing for video games as the success of the final product directly depends on it being used (played and enjoyed) by users (players). Stories can be a way to keep players at the centre of the game development process.

Game development includes many disciplines, each with its own interpretations, rhetoric and formalities. Storytelling can bridge these disciplines and help in building a shared vision and interpretation, by providing examples and a common vocabulary for everyone in a development team. We can use stories to gather, share and distribute information about players, tasks and goals (e.g., their motivation for playing a game). Stories can be a powerful tool in game development for encouraging collaboration and innovation of new design ideas across the whole design team (from programmers to publishers).

BioSt make use of storytelling (or storyboarding using graph representation) to point out user test findings and their impact on player experience. Storyboards could visualise the impact of an improved design on players' feelings. For example, we could use the storyboards to visualise positive and negative player emotions during gameplay as well as player engagement. Matching user test reports to these observations can provide a powerful overview of game levels and help uncover game design weaknesses. Storyboarding could also be seen as a powerful tool for triangulating or combining different data sources, bringing together the power of quantitative data as well as the depth of insight gained from qualitative inquiry and observation. The BioSt approach aims to help game user researchers to visualise game design intentions, player experience reports, and physiological responses. The BioSt graph is drawn based on the integration of physiological data annotated by player-defined experience points in the game.

The limitations of stories (if they are not data-supported) can be that they are a personal and subjective account told from a consumer's perspective. Therefore, recording and assessing experiences can become fairly intangible with subjective narrative accounts. However, stories can become useful when we generate player stories based on data collected during user test sessions. For example, these data may include player comments, observational notes, gameplay metrics and biometrics. Analysing these large-scale or high-resolution player data (e.g., analytics or biometrics) can be daunting, and presenting results from these studies is often not straightforward. Using stories would help the understanding of the human aspects in these data. These data-supported stories would help game developers understand GUR reports better.

4.2.2 Datasets for BioSt Prototypes

In BioSt prototypes the player story graph is drawn based on data gathered during the user test session, namely:

(1) Player's biometric responses

(2) Player's post-session interviews to explain 'why' the change in their signal occurred

(3) Player's self-drawn diagrams of their gameplay experience and

(4) Manually coded player gameplay behaviour (or context) based on observation.

These datasets were explained in Chapter 3, the next section briefly revisits them and explain how these datasets have been refined during the prototyping phases to better fit for the BioSt tool, the needs of the game development cycle and the fast turn around for the GUR results.

The prototypes utilise measurements of arousal in players' GSR during user testing sessions. Specific events with higher impact on the player's feeling are identified as potential usability or UX issues, generating a log of issues for analysis. In the post-session interview immediately after the gameplay session, each player was asked to recall these specific moments and to inform the experimenter of their thoughts and most importantly 'why they felt that way', with the video footage available for playback if the player did not recall fully. This approach helps to identify not only the negative usability and user experience issues, but also the events in the game, which have a positive impact on player experience. These selected events are visualised by ups and downs in BioSt graphs, and they are annotated based on player's comments.

Player's self-drawn diagrams are used to capture a player's overall experience of each level. This has been used as extra data to support player's comments from post-session interviews. As explained in Chapter 3, players were asked to draw these diagrams immediately after they were finished with the gameplay. These diagrams also show that players can mostly recall details from the beginning and end of their gameplay session. Combining the player's drawing and biometrics with post-session interviews helps to address this recalling problem.

Moreover, to better represent an accurate reflection of the player's gameplay experience, data from observations of the player's gameplay behaviour are also included when generating the BioSt prototypes. This is aimed to explain 'why' the player experienced certain changes in their physiological states.

4.3 First Case Study[1]

4.3.1 Setting

An unreleased commercial single player first person shooter console game was subjected to single player user testing for approximately one hour per player. At the request of the development team, 9 players were recruited from two informally identified demographic groups: 5 self-identified mainstream gamers (more casual gamers who play occasionally) and 4 core gamers (experienced gamers who play FPS games frequently). Only a portion of the game was completed to a level of quality indicative of the final product, and only these sections were tested. Testing was conducted over 3 days in a GUR lab. Participants played the game on an Xbox 360 connected to a HD (high-definition) television. Video cameras recording the player sitting positions, biometrics kit capturing the player's GSR and real-time footage from the game console was simultaneously streamed to the observation room next door. All feeds were composited together on a single display and recorded for later analysis. The games producer, and GURs monitored the participants' play from the observation room. The GURs had spent some time familiarising themselves with the game before the test sessions, and the producer was able to identify when players were not playing the game as intended.

Each session was conducted by one player at a time; before the session started, the players received a brief explanation on the session, signed the consent form and NDAs, GRS sensors were fitted and they were asked to relax for few minutes in order for the signals to be tested and stabilised. They then played the game for about one hour. After the gameplay, GSR sensors were detached, and the players were asked to draw their experience diagram. The GURs then interviewed the players (e.g., asking about their likes and dislikes), and after this short interview the players were asked to review their gameplay video on the GSR selected events; the GURs noted player's verbal explanation on those selected events.

Following the user testing, a text report was produced by the GURs who ran the session. The report listed gameplay issues encountered by each participant, as well as some additional design recommendations. As for the focus of this thesis, a BioSt was also drawn manually (by the thesis author), visualising issues identified and reported in the text report. At this stage, the interest was to evaluate how game developers interpret and use the BioSt. The next section provides a brief overview of gameplay issues identified from the user testing, followed by an explanation of the first BioSt prototype.

---

1 The three case studies detailed in this chapter were conducted on commercial under-development titles. Due to the NDA (non-disclosure agreement) requirements the names of the titles and any recognisable information has been removed. For the purpose of this thesis, the focus of these case studies was to evaluate the BioSt prototypes in real-case scenarios and not on the quality of identified gameplay issues as a result of user testing.

4.3.2 Overview of User Test Findings

Although the focus of this chapter is not on quantifying the gameplay issues uncovered during the user testing, the overall identified gameplay issues are briefly mentioned for each case study to give context to BioSt prototypes.

Results from the user testing show that both casual and core gamers reported similar experiences of the game, both groups found the enemies AI (artificial intelligence) were quite easy to beat and players often set their own challenges (e.g. only use headshots). Hence, players played the game like any other FPS and didn't feel the need to use unique features of the game.

In the build tested, there were several events when gameplay became frustrating and led to boredom, these sections were consistent for both core and casual gamers. Overall 61 gameplay issues were identified.

Key Positives:

- Open world combat - the ability for players to take their own route enhanced the tactical feel.
- Enemy variety - different types of enemies helped make the game feel unique.
- Atmosphere - audio and music seems effective at inducing excitement or tension.
- Pacing - regular periods of combat (once past the initial tutorial section).

Key Issues:

- Feature X - players did not use the feature X and ended up treating the game like a standard FPS.
- Clear objectives - players became frustrated when they were not sure where to navigate or their objectives were not clear.
- Tutorial - the tutorial covered the basic movement mechanics, but the unique features of the game (such as feature X) were not thoroughly introduced.
- Level readability - several key areas of level design were identified as causing frustration, players could not determine where to go or what to do.

4.3.3 Biometric Storyboards First Prototype

Following the user test sessions, a BioSt was created of each level for two random players. Figure 4-5 shows an example of BioSt's first prototype. Each vertical line is one minute of gameplay, peaks in the graph line can be positive or negative (e.g. enjoyment or frustration), comments have been coloured in green or red respectively to help identify the player reaction. For example, if the word 'combat' is written in green and is associated with a peak in the graph then this would indicate the player was aroused by fighting. However, if 'combat' is written in

red, then this would indicate that the player was not engaged or frustrated during combat. Therefore, positive comments are in green and negative are in red. In-game screenshots are shown beneath the graph as visual cues of where the player is in-game (the screenshots are also removed to the game developer's NDA).

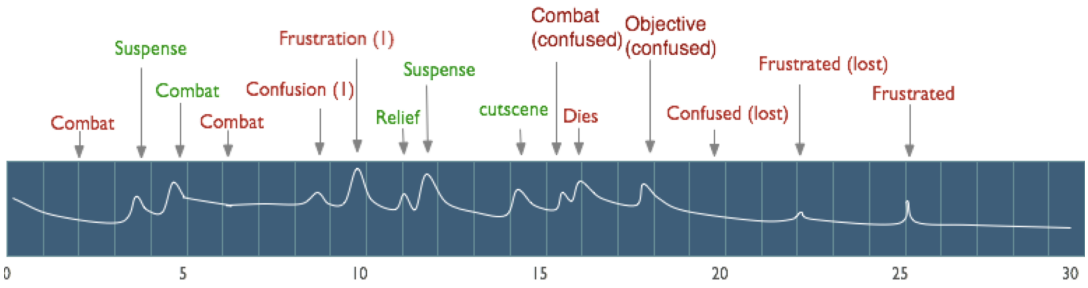


Figure 4-5 Example of BioSt's first prototype

This first prototype was divided by time. The feedback from the game development team suggested that this design of BioSt was difficult to compare between players. Time is not always meaningful for some games, and segments (or thematic areas) were considered more representative, as used in the second prototype. The game development team also commented that the use of in-game screenshots did not provide any useful information, as the team is familiar with game/level events and there was no need for visual cues of where the player was in the level.

4.4 Second Case study

The second prototype of BioSt was refined based on the feedback from the game's producer. The feedback suggested that the first prototype of BioSt was difficult to compare between players. In order to answer this main requirement, the following changes have been made to the BioSt prototype drawn after the second case study:

1) Each level was divided into thematic areas (segments), this would make the key sections easier to compare; it also shows the time it took the player to complete that area.

2) Green or red dots shows the positive or negative experience.

3) Annotation comments were also replaced by indexing numbers, where they pointed to explanation of each peak.

4.4.1 Setting

The second case study was conducted under similar conditions and with an updated build of the game to the first case study, but with a different set of 9 participants. As requested by the game developers, users were recruited with a variety of gameplay experience, with each playing the under-development game for approximately 1.5 hours. Similar to the first case study, players

received a brief explanation on the session, signed the consent form and NDAs, GRS sensors were fitted and they were asked to relax for few minutes in order for the signals to be tested and stabilised.

After the gameplay, GSR sensors were detached and the players were asked to draw their self-assessment diagram. This has been used to compare what the player perceived with what they actually experienced. The GURs then interviewed the players and they were asked to review their gameplay video on the GSR selected events; the GURs noted the players' verbal explanation of those selected events.

Following the user testing sessions, a text report was produced by the GURs who ran the sessions. The report listed the gameplay issues encountered by each participant. A BioSt visualisation was created, covering similar issues indicated in the text report. Although the focus of this chapter is not on quantifying the gameplay issues uncovered during the user testing, the identified gameplay issues are briefly mentioned in the next section, just to give context to BioSt prototypes.

4.4.2 Overview of User Test Findings

The focus of the user testing sessions was to evaluate how users understand, interact and their initial experience with the latest build of the under-development game.

Overall, 17 gameplay issues were identified; 1 issue was categorised as severe, indicating a critical barrier to gameplay, 11 issues were categorised as major barriers to understanding gameplay or cause of user error. 5 remaining issues were a hinder to gameplay, or impacted on the overall experience.

Key positives:

- The more skilled players enjoyed the multiple routes that can be taken to reach an objective.
- Many players commented that the visuals were high quality.

Key issues:

- Players did not use the Feature X, even when they probably needed to, e.g. heavy combat. Most treated the game like any other generic FPS.
- Many players commented that they wanted more combat scenes.
- Key gameplay concepts were not introduced clearly which led to player confusion.
- Navigation was still an issue in many places, particularly on the 2nd level.

4.4.3 Biometric Storyboards Second Prototype

In the second prototype, the levels were divided into segments, for example the first level was divided into 13 game areas. These areas can be seen across the top of the storyboard. The other key features can be seen on the sample below:

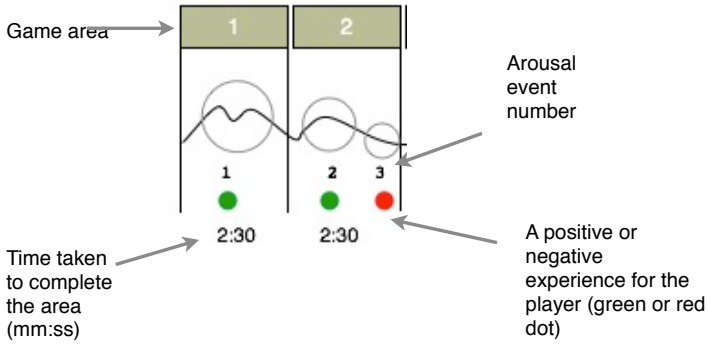


Figure 4-6 Key features of the BioSt second prototype

Figure 4-7 shows the second prototype of the BioSt, created from the second case study. The following are the main iterations in this prototype: 1) each level divided into thematic areas, this aimed to make the key sections easier to compare; to show a time it took for a player to complete that area. It also made it easier for the game designers to see where the issues were located exactly. 2) A brief text explanation on player's experience story (with reference to arousal events) added to the BioSt. 3) Green or red dots showed events with positive or negative experience (concluded from players' drawing and post session interviews).

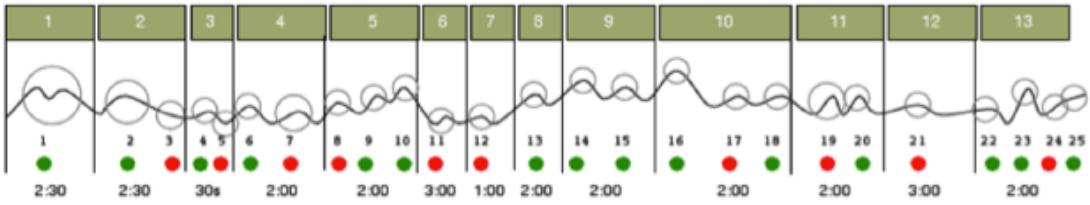


Figure 4-7 Biometric Storyboards second prototype. Examples of arousal explanation: 1: player interacts with objects. "Nice being able to move these objects." 2: first open space. 3: too much text instruction bored the player. Example of Player gameplay story: "The biometrics measures and player diagram shows that the player enjoyed the beginning of the first level. He enjoyed being able to interact with objects (1,15). His biometrics shows a positive reaction in moving toward different environments in this level (2, 14, 22)…"

Examples of arousal explanation: 1: player interacts with objects. "Nice being able to move these objects." 2: first open space. 3: too much text instruction bored the player.

Example of Player gameplay story: "The GSR measures, interviews and player diagram show that the player had positive experience from the beginning of the first level. He commented that he enjoyed being able to interact with objects (1,15). His GSR shows a reaction in moving toward different environments in this level (2, 14, 22)…"

Feedback from the game developers suggested that the graph should go down (negative gradient) to indicate negative player experiences in order to better represent the emotional change, and to better draw attention to and isolate the negative experiences. Secondly, they reported difficulty

in pinpointing the exact moments highlighted by the red/green dots, which were key to providing context and to establish cause and effect.

4.5 Third Case Study

To evaluate the third prototype of BioSt and explore the possibility to use this approach on different types of video game, a third case study was conducted on an unreleased commercial sandbox (free-roaming) game. The third prototype (Figure 4-8) aimed to make the diagram easier to read and couples behaviour (the text along the bottom) with the associated player experience. This prototype of BioSt was refined based on further comments from the game developers (from the second case study). The following changes have been made to the BioSt prototype created for the third case study:

1) The experience graph to visually represent a negative gradient to indicates negative player experiences and better represents an emotional change (although either negative or positive reactions to micro-events may result a peak in GSR readings).

2) Red/green boxes to place on the graph in order to draw attention and pinpoint the exact place of issues, which are key to providing context.

4.5.1 Setting

The third case study was conducted on an AAA under-development commercial sandbox (free-roaming) series franchise game with similar conditions to the first and second case studies. As requested by the game publisher, 8 users were recruited; 4 of them familiar with the game series franchise, and 4 of them new to the game series franchise, with each playing the under-development game for approximately 1.5 hours. Similar to the two previous case studies, players received a brief explanation on the session, signed the consent form and NDAs, GRS sensors were fitted and they were asked to relax for a few minutes in order for the GSR to be tested and stabilised. After the gameplay, the players were interviewed and were asked to draw their experience diagram.

Similar to the two other case studies, a text report was produced by the GURs who ran the session. The report covered gameplay issues, as well as some additional recommendations. The third prototype of BioSt was also created, covering similar issues as noted in the text report. The next section provides a brief overview of gameplay issues identified from the user testing, followed by an explanation of the third BioSt prototype.

4.5.2 Overview of User Test Findings

As requested by the game development team, the goal of the user test sessions was to explore if the players enjoyed the game, what led to their enjoyment, and if the game experience was different for players new to the game series franchise.

The report covered the following key findings:

- New powers are not introduced in a 'best practice' approach
- Combat can be disappointing
- Side missions are not satisfying
- Game mechanics are not clear
- Backstory needs better communication

4.5.3 Biometric Storyboards Third Prototype

Figure 4-8 shows the third prototype of BioSt. As shown in the figure the level was divided into segments, related segments to each mission were colour coded. The drawing of the experience graph aimed to better indicate negative player experiences and to draw attention to and isolate the negative experiences. Red and green boxes (indication of the player's positive or negative comments) were also placed on the graph with an associated brief explanation. As requested by the game development team, this prototype also included the player's rating of each mission and an indication of the player's avatar death, shown by "x" above the graph.

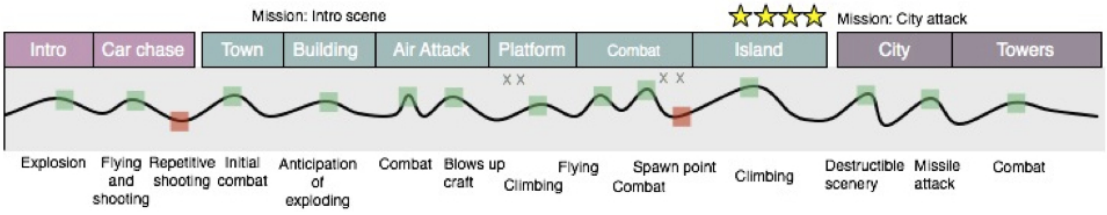


Figure 4-8 Biometric Storyboards third prototype (Due to the game developer's NDA the segments names and some explanation information have been anonymised)

The game publisher's feedback suggested that this visualisation provides them with a quick overview of the areas delivering positive and negative experiences. They also mentioned that they have used BioSt to discuss and communicate game design decisions.

4.6 Prototypes Evaluation[2]

After having three prototypes on the BioSt visualisation idea, the next step was to evaluate these prototypes with more game developers. This section reports further evaluation allowed by showing the three prototypes to game developers who had not seen BioSt before. The BioSt prototypes were presented to six game developers and they were interviewed about the advantages and disadvantages of this technique.

4.6.1 Method

Semi-structured interviews were conducted in order to evaluate the three iterations and better understand the target user requirements for future development of the technique. Six game

[2] The study presented in this section was published as a work-in-progress paper at CHI 2012 (see publication list P5).

development professionals (P1: Lead Designer, P2: Creative Director, P3: Designer, P4: Programmer, P5: Animation Designer, P6: Game Director) from a midsize UK game design studio were interviewed for this study. Since the prototypes were developed during GUR studies and with involvement of GUR professionals, it is important to understand how game developers would perceive these representations of data. None of the interviewees had seen BioSt before. Each interview took about 30 minutes. Before the interview the participants signed a consent form. The written notes from the interviews were analysed and then categorised into themes (Figure 4-9).

Figure 4-9 Categorised interview results

Each interview started with the interviewee's thoughts on user testing and user test reports. For example, they were asked about their overall experience with user testing (UT), what they were hoping to get from the report, about its format and how findings were presented (communication). After this opening discussion on UT and reports, the interviewees looked at all three prototypes of BioSt and discussed advantages and disadvantages of this technique and each prototype. In order to reduce any potential positive biases, the participants were made aware of the work-in-progress nature of these prototypes and were encouraged to provide feedback on how the work can be developed in a way that fits into the game development cycles.

4.6.2 Results

All interviewees had read UT reports before and some of them were involved in conducting UTs in their studios. The main values of UT sessions for P6 are to see: (1) areas of frustration, (2) areas that are difficult to pass (blockers), (3) if the players are having fun, (4) if players understand the game and are using all the game features. They mentioned a need for more visualised data in UT reports. An ideal report would be a process to capture a massive UT data and report it in a way that is easy to make sense of. For example P2 mentioned in their previous title (a racing game) that they collected game metrics to generate a crash heatmap of each track. He added: "*from heatmaps we could see the crashes, but we know they can lead to different experience. Some of them lead to enjoyment and some lead to frustration, the heatmaps won't*

show this difference, [...] BioSt is somewhere between only seeing the heatmaps and talking to the actual players." [P2] The interviewees suggested the text reports usually cover most of the information they need but they also mentioned: "*for some issues you won't feel the text report can put them in a right context and time line. For example when interpreting from a report there is no way to see the change of pace and enjoyment*." [P6]

The interviews suggested the following advantages and possible improvements for BioSt prototypes:

- At a glance summary
- Objective credibility
- Location and prioritising of gameplay issues
- Identifying a problem/suggesting a solution
- Clarity/simplicity
- Facilitates the discussion
- Trust/convincing
- Comparison to intended experience

The report summary is the section they always read and found most useful. All of the interviewees think BioSt shows an at a glance ***summary*** of a level.

P6 believes: "*using biometrics lends more credibility to BioSt. It gives the perception that it includes data that goes beyond what players say but what they feel, stuff they don't realise to vocalise it.*" This would suggest using biometric measures as one of the data sets for creating BioSt can offer perceived ***credibility*** from game developers point of view.

Location of issues in each level is an important factor to priorities what to fix. BioSt can show exactly where issues occurred. "*BioSt allows me to see where my good and worst parts are, it helps me to prioritise what to fix*." [P6], "*this* [BioSt] *shows me pretty much negative experience happening at the beginning of the game/level, it's concerning*." [P1]

The developers do not want UT reports to ***suggest a solution*** on how to fix an issue. "*I just need to know where the problems are and how much of the problem it was.*" [P5]. For them the ideal report would be a combination of text to explain what the problem is, illustrated with short gameplay videos of a player experiencing the problem. BioSt can show them where the problems are and also, by visualising the relationship between issues, it may assist developers to come up with a possible solution.

Clarity: As a visualising tool it is critical that the developers are able to correctly interpret the data from BioSt. P5 said: "*this is very clear, easy to see the different sections. It is difficult to contrive this to anything else.*"

The interviewees were from different positions in game studios, yet they all felt that BioSt would be helpful for them. They also mentioned BioSt would provide useful information to publishers and studio executives. P6 mentioned "*the issues are usually not small enough to be actionable by single person but this* [BioSt] *can be used with the whole team* ***to facilitate the discussion*** *over a level design*." P3, P4 and P5 mentioned that they want to see more data such as players' comments and gameplay video in each indicated events. P3 added: "*my view is from a designer's perspective, where we are eager to go into details, like user comments, to see what this guy said about this bit*."

Trust is a vital matter for UT reports, "*if the designers do not trust the data the problems will stay*." [P2]. The interviewees suggested that the most convincing case is when the designers personally attend UT sessions and have a face-to-face conversation with players or watch the gameplay video. As for the content of a UT report, P3 said: "*It will be wrong if we ignore any statements, but we act on it if many people say same thing is wrong*." Specific to BioSt, P2 mentioned: "*there are two ways I can trust it, one is for me to totally understand how it is generated, or to see enough correlation between the results*." This is explored further in the discussion section.

Comparison to intended experience: The interviewees mentioned they use some sort of storyboards depending on the game they are developing. For example P2 explained that for their previous game they draw a graph of intended emotional states for players. BioSt could make it easier for developers to be able to compare the player's experience with the experience they intended to design for. "*If BioSt can show the accurate match to our intentional graph that would be a fantastic tool*."

4.6.3 Key Findings

This prototype evaluation study aimed to improve on the following areas for the next step of the work:

Overall prototypes: after seeing all the three iterations the developers overall feedback on them was:

- All interviewees preferred 3rd prototype.
- Positive feedback for adding level areas in prototype 2 and 3.
- Positive feedback for having graph annotation and area descriptions in 3rd prototype.

- Negative comments on 2nd prototype as they experienced problems with finding area and arousal explanations. This was fixed for the 3rd iteration.
- Negative feedback for removing time player spent in each area from iteration 3.

Composite graph instead of individual: The current design of BioSt visualises how each individual player experienced a game. Based on the interview results this can lead to two problems: 1) Too many individual graphs for developers to look at and 2) Showing how one individual player experiences a game does not convince developers to act on the issues. In order to improve these we need to generalise the individual graphs into a composite graph, showing the correlation of results among players. The results suggested that the BioSt would be a useful tool if it shows a reasonable correlation between the results of individual players.

Severity: The developers want a tool to help them to prioritise the issues to fix. While BioSt facilitates this in some ways (e.g., location of issue) it could also indicate the severity of each issue. For example this can be achieved by measuring the mean value of GSR arousal among participants at the specific event.

Interactive: The developers want BioSt to be interactive. For example, enabling a mouseover on each point to get the description or to click on each area to zoom in and see comments from different players, or to watch a clip of their gameplay video indicating the specific problem the developer is looking at.

Graph comparison: Developers want to see their intended experience graph in the BioSt, or a graph that can show the difference between intended and actual experience. This would be possible if the developers would be willing to work closely with user researchers and provide them with their intended emotional graph.

Measurement of different experience: The interviewees suggested the current design just shows green and red points, as high and low arousal experiences. Since these do not actually depict emotional valence, adding other sensors (such as EMG, EEG) would allow approximating wider types of experience.

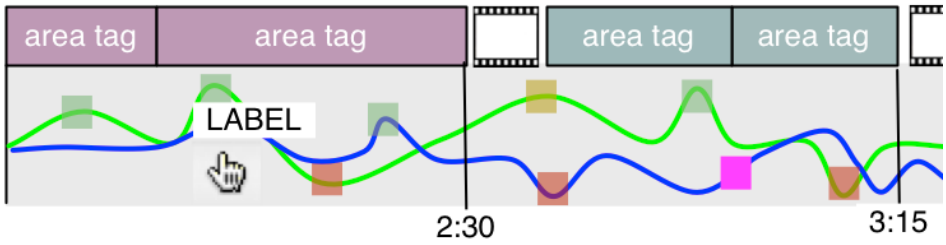


Figure 4-10 Possible next iteration of BioSt

Figure 4-10 shows an example design for a potential next iteration to the BioSt prototype. It is possible to generate two experience graphs to facilitate the comparison between them as explained above. For example, the two graphs can be a comparison between core and casual players, or designer-intended and an actual experience graph or collocated players. Possible different experience types may be visualised by using colour boxes. For example, players' reported gameplay experience might be coded and mapped to different colours. A BioSt can also be interactive, for example, to enable a mouseover, or a click for zooming-in or displaying more data. These will be discussed further in the next chapter where the focus is on the BioSt tool development.

4.7 Discussion

The BioSt visualisation prototypes presented in this chapter use GSR to show the potential to make difficult-to-interpret physiological GUR data more accessible to a wider game industry audience. Some of the key strengths of the BioSt approach identified from 3 case studies and the prototypes' evaluation are:

***Correlation between physiological responses and gameplay events*:** Visualisations of players' physiological responses and gameplay events are helpful to understand and explore correlations of physiological responses and the corresponding gameplay events. For example, by mapping a player's GSR, post-session interview comments and the game events, BioSt shows potential for being used to explore the behaviour of several players during a single game event. Understanding how players are motivated to perform particular tasks in gameplay environments is a vital tool for game designers.

Comparison of players' behaviour: Once a series of these BioSt are created, they could be used to compare the gameplay journeys of different players and spot key trends in gameplay behaviour. Further studies may show that a player's background profile and psychographics can reflect a regular pattern of behaviour and subsequent enjoyment in their corresponding BioSt. For example, the potential to visualise the experience of co-located players, such as collocated player vs. non-player (observer).

Whole session overview: By visualising the whole gameplay session, BioSt were able to provide an efficient overview across all events, levels, and missions, enabling the developers to quickly scan for key elements in level design, player performance and player emotions.

Verifying the intended design decisions: the case studies showed that the game developers tend to exploit BioSt to compare how players experienced the game events with what the designer had originally intended. Suitably equipped to understand the effectiveness of the design evaluation process, these developers were able to verify the success of their game design

environment, and judge whether the intended game experience matched the actual player experience. By providing a tool to facilitate comparing between intended and perceived design, BioSt (with further development) could be used as an experience design tool.

Simplicity: BioSt was prototyped based on the demands of game developers to deliver a tool that is easy to understand and interpret with an immediately apparent benefit.

User-centred Design: Understanding of the game development process and the relevant needs in the working environment has helped to design visualisations which closely match the requirements and language of the target users, and the subsequent level of detail necessary for the task.

Familiarity: Game developers and producers are familiar with various data representation techniques and visualisations of game metrics. Similarity between BioSt and these existing models helped to support communication with and between developers and effectively increased the acceptance of new tools.

Support collaboration: By visualising player experience issues in gameplays, BioSt may facilitate collaboration between game user researchers, game designers, game developers and producers in order to discuss and fix identified issues. Comments from case studies suggested that producers and designers were able to more effectively discuss design strategy using BioSt as evidence for player behaviour.

4.8 Conclusion

Overall, the recommendations generated from the three prototypes, together with a user-centred design (UCD) process, has helped to design an approach that has provided game developers with an increased understanding and enhanced communication, leading to a better understanding of player's gameplay experience. Development of the BioSt technique was based on (a) simplicity in visualisation, (b) strong user involvement throughout the entire design process, and (c) an integration of the target users' existing tools (storyboards have been widely used in the video game and movie industries). A focus on simple visualisation and UCD helped to iterate the earlier prototypes to the current iteration.

Game events (or game segments) and emotions resulting from those events are two of many important aspects for creating a great player experience. BioSt aims to visualise events in gameplay where player's actions or behaviours lead to changes in their emotional states. Hence, it enables GUR professionals to have an accessible and fast way of analysing and triangulating physiological data, player reports and game design intentions.

4.9 Summary

Storyboarding can help to visualise different player data in one graphic representation. This can help GUR and game development teams to achieve a shared view on critical game design events. BioSt are not only a powerful tool to explain game design problems but also provide a way to discuss their solutions. They can help the whole team to visualise the design problems, the potential solutions, and gameplay areas that need improvement.

Creating data-driven storyboards supports design arguments, so that the game designers can see how players would experience their intended designs. These storyboards provide an analytical connection between players and game designers.

The BioSt technique aimed to generate biometric-based visualisations which provide better support to problem-solving and communication, greater insight into player gameplay experience, and better fits into the work process of video game development than traditional HCI user research methods. So far the results from the three prototypes discussed in this chapter and their evaluations show positive feedback on how BioSt have helped developers to gain a better understanding of how players interact with their game, ultimately enhancing their ability to effectively optimise the experience of the final release.

Chapter 5 will look at further development of BioSt, by creating an application that facilitates the collection and visualisation of player data from user testing sessions, in order to provide a formal approach to create BioSt more effectively and efficiently.

5 Biometric Storyboards: The Tool

This chapter provides a detailed explanation on how the BioSt software tool was developed, including which data to capture, analysis of the physiological measures and how the tool maps these measures into game events and players' self-report of their experience. After discussing the first and second phases in the design process in Chapter 4, this chapter discusses BioSt's tool feature list (based on the evaluation results) and provide in-depth development details on the BioSt visualisation tool.

5.1 Introduction

The main contribution of this thesis is introducing the BioSt method and hence the application developed to facilitate creating BioSt. This is an important contribution to the GUR field because it allows faster analysis of highly complex and interrelated mixed datasets from quantitative and qualitative sources, suited for the short turnaround cycles in game development companies.

Games user research traditionally relies substantially on self-report measures and behavioural observation to investigate player experience within video games. Novel approaches, such as game metrics and physiological signal recording during gameplay, have been introduced to provide more objective and covert measures of player behaviour without interrupting the gameplay experience. Affective physiological player evaluation in GUR involves measuring the electric potential on the surface of the skin to make inferences about muscular activity, brainwaves, arousal, and general emotional activity (Drachen, Nacke, Yannakakis, & Pedersen, 2010a; Nacke & Lindley, 2008). One problem that inhibits a broad industry adoption of physiological measures is that – due to their high temporal resolution – they produce large amounts of data that are hard to analyse within the short iterative development cycles of game companies, this is in addition to required user researchers that are trained and have expertise to conduct physiological evaluation. If applied, GUR teams use physiological measures in conjunction with subjective measures such as player interviews or questionnaires, because the physiological data alone is often not meaningful enough to interpret a gaming situation.

The goal of BioSt is to present a GUR method to facilitate the analysis of physiological data and provide qualitative annotation capabilities that would allow players to comment on the physiological data gathered during a gameplay session. Together with a designer-drawn player

experience graph (see section 5.4.1), this allows a triangulation between designer intention, player experience, and player physiological responses, to create and visualise a meaningful data set for improving game design. Physiological measurements of players have been used as well as their post-gameplay comments during game events, for providing a meaningful relationship between changes in player's affective state and their gameplay experience.

Since the volume of physiological data is hard to explore because of the large amounts of data collected, a visualisation approach has been adapted that allows a comparison between participant-created storyboard visualisations (from their post-gameplay comments) and physiological data visualisations. In addition, BioSt use this visual narrative to allow easier interpretation of the mixed method data. Making meaningful connections between self-reported player experience narrative and measured arousal is an important step for increasing acceptance of physiological measures in the GUR field.

5.2 Interaction Steps

The developers' feedback received from prototypes evaluations (Chapter 4) fundamentally guided necessary design features for the BioSt tool. In order to provide a meaningful visualisation of GUR data, the tool was designed to provide systematic steps to gather GUR data, as well as to facilitate the analysis of physiological data and provide qualitative annotation capabilities that would allow players to comment on the physiological data gathered during a gameplay session.

Figure 5-1 shows steps that can be followed to generate BioSt report:

1) Game designers draw their intended player experience graph using the tool (section 5.4.1).
2) Player data (physiological responses and post-gameplay comments) collected using the tool (section 5.3).
3) GURs draw aggregated PX graph by viewing designers' intended PX graph, the player's physiological response along side with their comments from the post-gameplay interview (section 5.4.3).
4) The tool represents all these data in one report page. This enable game designers and GURs to easily compare the intended experience graph with the actual experience graph drawn by GURs after reviewing player data (see page 99).

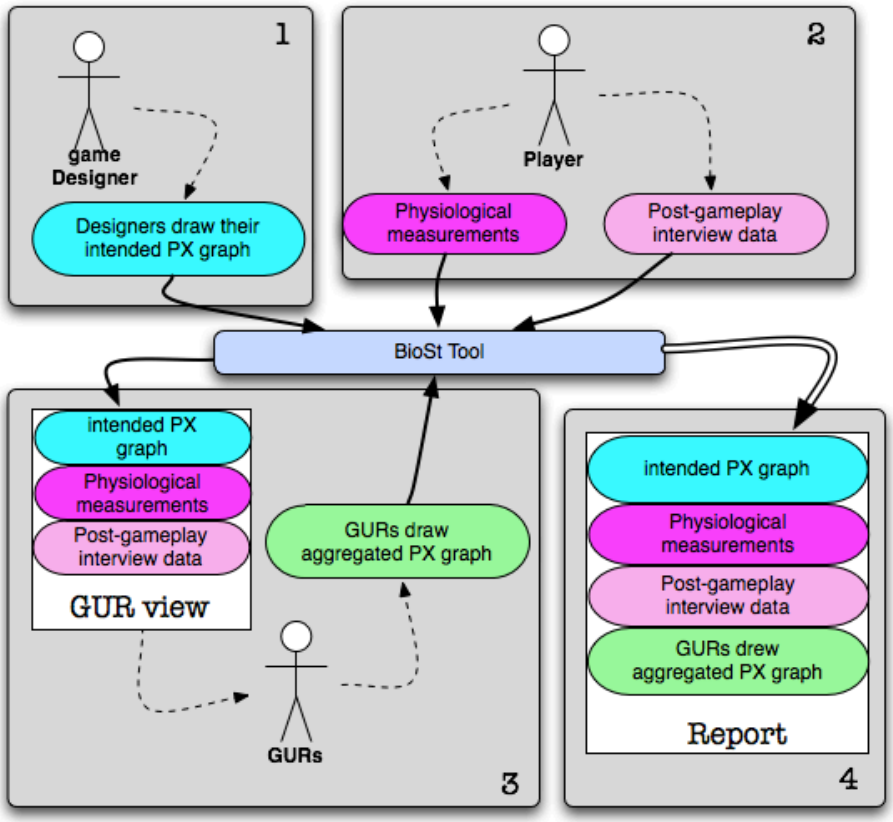


Figure 5-1 BioSt tool system design

5.3 Measuring Physiological Data[1]

Based on the prototype evaluation, it was decided to utilise GSR (indicating changes in player's arousal level) and facial EMG at zygomaticus (smiling) and corrugator (frowning) muscle locations as an indication of change in player's emotional valence level. GSR level is one of the easiest physiological measures to apply and analyse. Together with the similarly easy-to-use facial EMG (for example, EMG uses 5 sensors in comparison to EEG which uses 16 to 32 sensors), these two physiological signals provide relevant data for player engagement and the ease of application necessary in the games industry.

The NeXuS-10 MKII device was used to record physiological measurements; recording software was a custom C++ application using the NeXuS SDK to collect raw data from the device and display the recording timestamp on the computer screen (timestamp display enabled mapping between the physiological data and in-game events).

[1] Programming the BioSt tool and physiological measurement software were completed in collaboration with two undergraduate research assistants (J. Gregory and M. Beig) under the guidance of the thesis author. All other aspects of the application development and features were conducted as part of this Ph.D. research.

Once run, the recording program asks for a participant ID (to name the corresponding data files) and aims to connect to the NeXus. On successful initialization, the raw ascii data file will be created (and labelled using participant ID), and the thread will be signalled to begin recording the raw data with its timestamp. At this point the raw data is being recorded into an array and written to file as the program runs (at 60Hz). For every recorded line of raw data, the counter variable is incremented by one. As the data is recorded the raw value of each channel in use (in this case channels C, D and E as default channels for GSR and EMG readings on the NeXuS-10 MKII) can be analysed and compared to the previous minimum (min) and maximum (max) values, where the starting min and max are the highest and lowest values a float can hold on the hardware respectively. On a safe exit, the program will save the raw ascii output file, and then run post-processing on the data and save the processed text output file.

5.3.1 Galvanic Skin Response

GSR is regulated by the production of sweat in the *eccrine* glands, where increased activity is associated with psychological arousal. This makes it an ideal physiological measure for analysing games, where exciting moments are likely to elicit high arousal and engagement in the game. In BioSt tool GSR is measured using passive sensors attached to the *medial phalanx* of the ring and little fingers on the player's left hand (Figure 5-2).

Figure 5-2 GSR Sensors

The raw data collected from NeXuS Auxiliary Channel (E) are in millivolts (mV). These values have to be converted to microsiemens (μS) to reflect measurements comparable with the existing research literature. The listing below shows a formula provided by MindMedia[2] :

GSR (kOhm) = output value (mV) * Amplification + Offset

SC (uS) = 1000/GSR

Amplification = 1.38

Offset = 4.3

[2] The NeXus-10 Mark II is made by MindMedia: www.mindmedia.nl

After the above signal conversions are completed, the next step is to normalise the GSR values for each user. Since arousal levels differ from person to person, absolute GSR values are not comparable among players and are therefore converted to relative values by normalising them using the conversion below, a standard GSR normalization formula (Mandryk, 2008). A sample MATLAB code is used to calculate the normalise value:

```
MaxGSR = max(GSR);

MinGSR = min(GSR);

GSR_Normalized = 100*(GSR - MinGSR)/(MaxGSR - MinGSR);
```

5.3.2 Facial EMG

EMG sensors measure electrical activation of muscle tissue, and facial EMG has been used in emotion detection (Cacioppo, Tassinary, & Berntson, 2007). *Zygomaticus major* (smiling) and *corrugator supercilii* (frowning) facial muscle activity is measured for BioSt tool using passive EMG sensors on the player's cheek, brow, and ear lobule (for ground sensor) shown in Figure 5-3.

Figure 5-3 Attachment of facial EMG sensors (Cacioppo et al., 2007)

The raw data is analysed to indicate muscle activation in corrugator and zygomaticus muscles. The similar analysis approach was followed as described by Hazlett (2008), where a muscle is consider significantly active if the signal was above a threshold value of total sample average (M) plus total standard deviation (SD). After calculating the threshold values for both muscles, thresholds were compared to each sampled EMG value. If the sample value was above the threshold, the measuring muscle was noted as active at that moment. This is stored as a Boolean value for each time step. If the current processed value is greater than or equal to the cut-off, then 1, or else 0, is recorded as true and false respectively. Below shows an example MATLAB code used for this calculation:

```
[a b] = size(EMG_C);
```

```
for i = 1:a

if EMG_C(i) >= THRESHOLD

   EMG1(i) = 1;

else

   EMG1(i) = 0;

end
```

5.3.3 Output Text File

The post-processing described in the above section generates a text data file that is then imported by the BioSt tool (see next section). The output text file contains the frame counter and processed data from the recording channels. Each frame counter is assigned with a normalised value of GSR data and a Boolean value for each of the EMG channels above the threshold. This is an example of the output text file structure:

GSR (E), EMG1 (C), EMG2 (D), COUNTER

56.277, 1, 0, 4610

56.881, 0, 0, 4670

57.458, 0, 0, 4730

5.4 The BioSt Tool

The tool was developed in the Unity game engine. This is to allow portability and interfacing game data collection in Unity games, as both could be potential areas for further development of the tool. The purpose of the BioSt tool is to visualise the data gathered from GSR, facial EMG measurements and user test sessions. The tool combines these data into a single view that can later be shown to game designers and compared to their intended player experience. GURs can also create an aggregated graph representing their findings from a number of players. In whole, the BioSt tool has three parts: (1) The intended player experience graph (2) The player's input view, and (3) The GUR view. This section explains each of these parts. The tool allows the designer to display aspects that they are looking to improve in their game. This is later compared with GSR, EMG and player's feedback for the GUR professional to create a composite graph that represents all of this information.

5.4.1 Designer's Intended Player Experience Graph

This part (see Figure 5-4) is where the game designers draw their expected player experience. The tool has functionality to represent areas of game including multiple events and segments. It

allows labelling these segments with an appropriate name and expected playtime, as well as key nodes or points in the player experience, (e.g., high to suggest an exciting experience or low to suggest a less exciting part – this also tolerates designers to reflect on their interpretation of intended player experience). This information is then saved for later use in the GUR view.

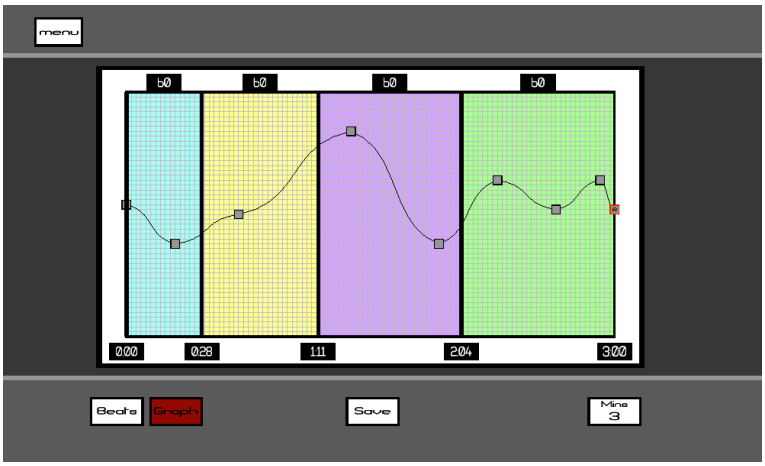


Figure 5-4 Screenshot of Designer's graph draw mode

5.4.2 Player's Input View

Iterated based on the results from studies S1 and S2 (Chapter 3), after a gameplay session, the player and researcher together review the gameplay video. The researcher types in the player's comments and naming of positive and negative experiences into the player input view of the tool, identified with a unique player's ID and event timestamp (this is so that the player's comments will synchronise with the physiological measurements) (see Figure 5-5).

Each player's data is viewed individually and is accessed by entering the ID. Once the unique player ID has been entered, the tool will either load a previous session or create a new account. To begin, the start and end frame of the recording should be entered into the tool. This allows the triangulation of input data with the biometrics.

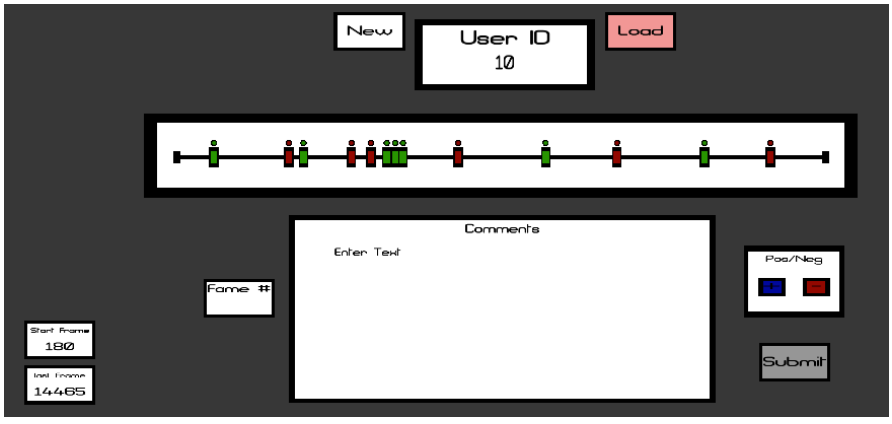


Figure 5-5 Player's Input screen

5.4.3 GUR View

The GUR view (Figure 5-6) is the main part of the tool, where the tool enables a link between the designer's *intended* player experience, and the player's *actual* physiological reaction to game events and self-reported comments. At the top of the GUR screen is always the designer's graph for quick comparisons between the data. The GURs' graph is located at the bottom and has the same functionality as the Designer Graph; allowing adding game segments, nodes and saving.

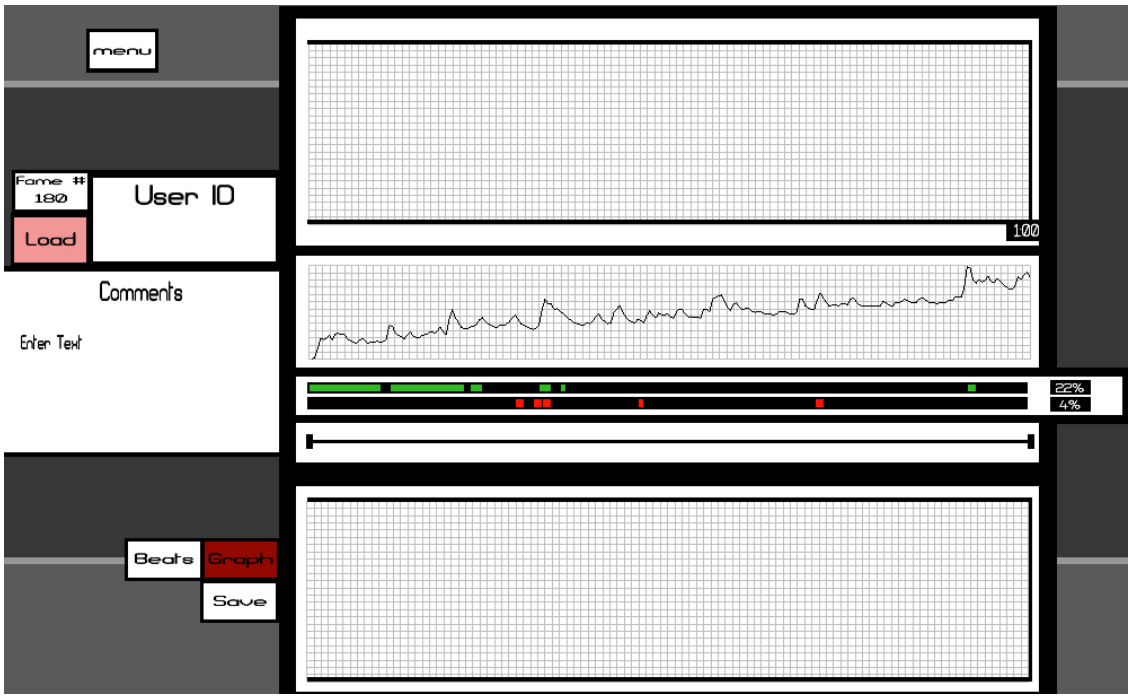


Figure 5-6 GUR screen view

The single player's data graph includes the player's physiological reaction to game events and their self-reported comments. Each player's data is viewed individually, showing normalised GSR, facial EMG measurements and the comments from the player. Figure 5-9 (next page) shows how the various elements of player data are represented along a timeline. It shows the normalised GSR graph, muscle activation bars (green: smiling muscle, red: frown muscle), index boxes for player's comments and experience descriptions, which are visible at mouse-over (blue: positive comments, red: negative comments). Game segments (i.e., game events) are also indicated by a coloured line and show the time a player spent in each beat.

Analysed EMG data is loaded from an input text file (automatically generated by physiological measurement software once the recoding is stopped) and the muscle activation is determined if the value is 1 or 0. If the value is 1 an active mark will be placed and its length is determined until a zero is read or reached the end of the file. Two separate entries are used to determine positive or negative readings; green boxes are activity markers in smiling muscle and red boxes

are activity markers for frown muscle. Normalised GSR is also read from the file as the y-value between 0-100. All of these readings are in sync using the same timestamp.

Further, GURs can aggregate these players' data as per the experience graph shown in Figure 5-10 (next page). A GUR aggregated player experience graph is located at the bottom of the GUR view, and has the same visual format as the designer's intended player experience graph.

The tool also has a feature to calculate overall positive or negative feeling over each level and shows this in percentage. This feature would be useful for example to compare an overall experience between conditions (such as different levels). (Figure 5-7)

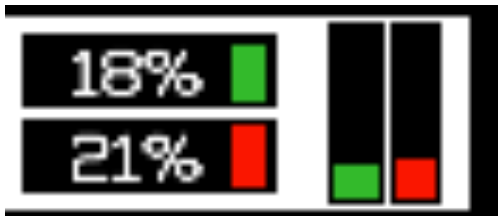


Figure 5-7 Percentage of smiling (green) and frown (red) muscles activity in one level

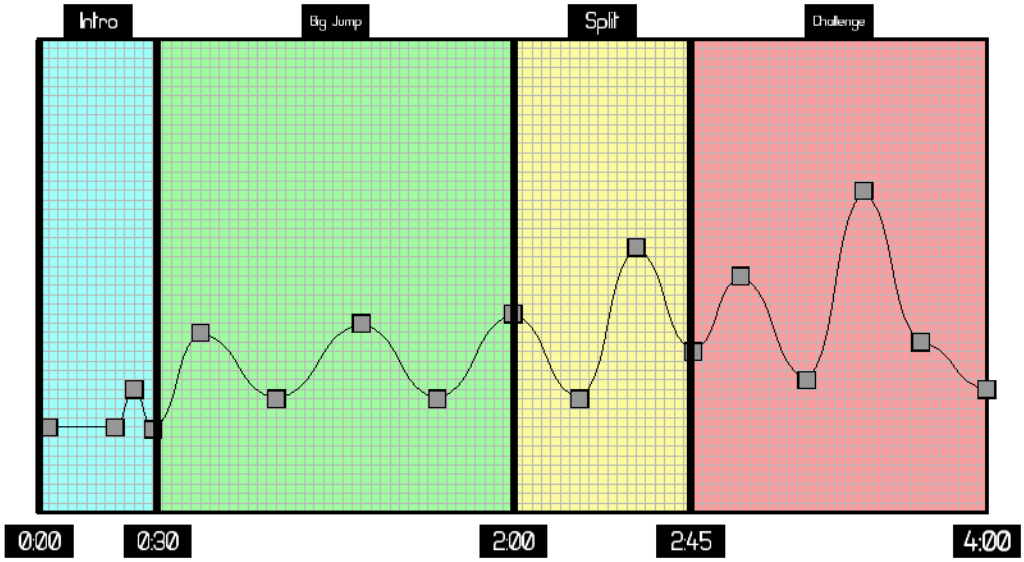


Figure 5-8 Intended player experience graph (representing what designers think exciting gameplay moments are) showing game segments, times and key nodes

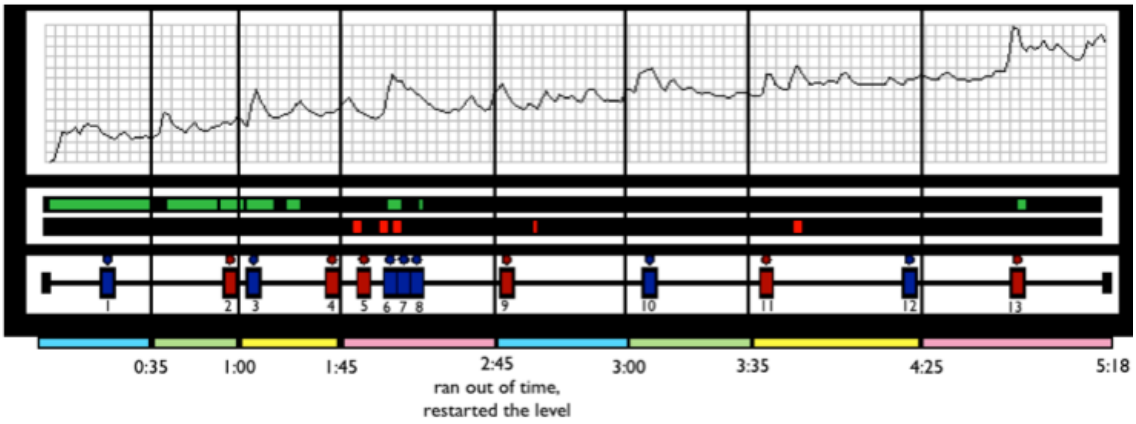


Figure 5-9 Single player's data graph in GUR view synced based on the frame counter timestamp

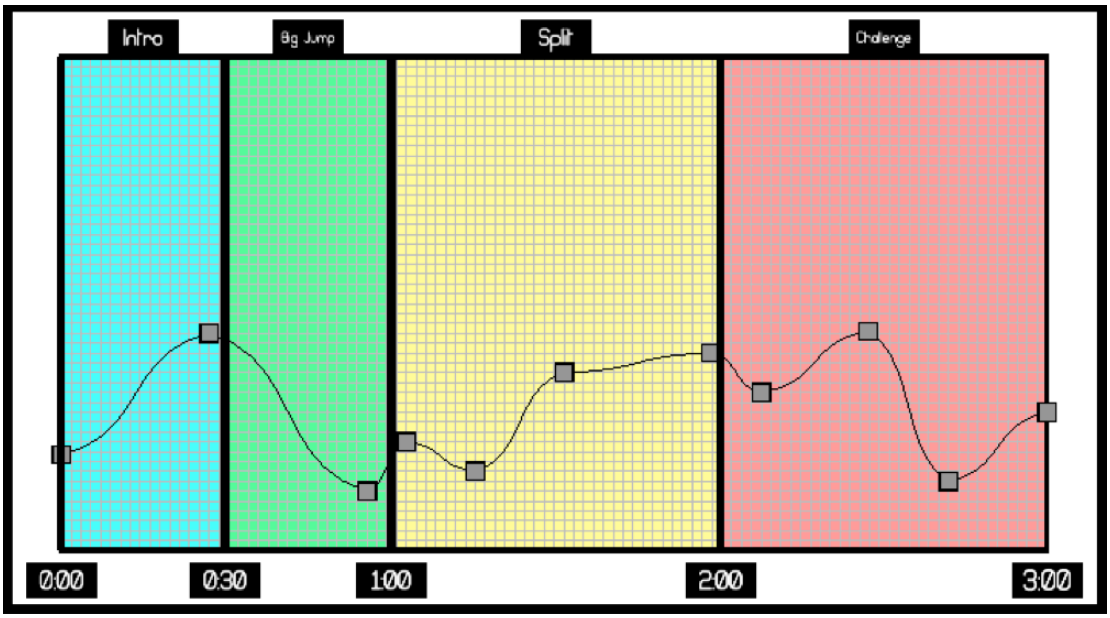


Figure 5-10 GUR aggregated player experience graph, indicating areas of difficulty and average time spent in each segment

5.5 Summary

This chapter focused on a detail explanation of BioSt tool and its development process. Hence, it covered the design of the tools interaction steps; EMG and GSR sensors' setup, placement and configuration; and interface for gathering and reporting the data. The next chapter details a study to evaluate the contribution of this approach in relation to other GUR approaches.

6 Evaluating Biometric Storyboards

Having developed the BioSt in iterative use and feedback from game developers, the focus of this chapter[1] is to evaluate the method and see if it achieves desired outcomes as well as to find out how it compares to other user research approaches. Hence, this chapter presents a study demonstrating how user research approaches and BioSt help designers create a better gameplay experience. In addition, this chapter shows that BioSt can higher gameplay quality than designing without any form of user test. The evaluations reported in this chapter support the idea that BioSt provides more focused feedback for game design improvement. This chapter also provides a discussion on BioSt contributions in GUR as well as the limitations of the approach and the study.

6.1 Introduction

This chapter evaluates the contribution of BioSt in a GUR cycle, from running user test sessions to fixing identified issues, by comparing the quality of three games: two designed with contrasting user tests (UT) methods and one designed with no UT method at all.

Previous work has demonstrated that physiological measures are suitable for evaluating user engagement with regard to the emotional component of their experience (Mandryk & Atkins, 2007; Nacke & Lindley, 2009). Hence, the game industry has shown interest in integrating these methods in game development evaluation (Ambinder, 2011; Zammitto, 2011). If utilising physiological measures, one of the major challenges for the game industry and researchers alike is tying together the results of physiological measures and player experience reports. To address this problem, the BioSt was developed, which combines designer intentions, user experience evaluations, and player reactions (physiological and observed) in a single UT report. Such a triangulation of data could provide valuable feedback for a game development team to optimise their game experience. However, since the field of physiological player evaluation is still emerging, there is a need to understand the usefulness of BioSt and its relative value in regards to other user testing methods and to prescriptive and intuitive game user research approaches.

1 The study presented in this chapter was published as a full paper at CHI 2013 (see publication list P1); the thesis author was the leading researcher for the paper, designed and conducted the study, wrote the UT reports and analysed and reported the results. The statistical tests and analysis were conducted by L.E.Nacke. The development of the game prototypes was completed by J. Gregory. This chapter is based on the published paper.

This chapter investigates the differences in game design between games developed using a classic UT (i.e., using observation, interview and questionnaires in order to conduct user tests, analyse and report gameplay issues), a BioSt UT (i.e., using BioSt tool to conduct user tests, analyse and report gameplay issues), and no UTs (i.e., designer expertise only). A recruited game programmer created three different versions of a game prototype based on the recommendations of game designers that used a classic UT, a BioSt UT, or no UTs, to create a list of design recommendations.

The study results support the GUR mind-set that UTs improve games considerably by showing that using either BioSt UT or Classic UT leads to games that are better designed and that rate more positively than games designed without any form of user testing. The key contribution of this chapter is providing evidence that games evaluated with classic UT or BioSt UT actually provide a higher quality product and a better gameplay experience. This implies that user test sessions with BioSt will provide more focused feedback for design optimisation, hence, result in higher perceived gameplay quality and provoke more subtle changes to game mechanics.

6.2 Related Work

The methods used in game user research have been extended and modified from existing HCI methodologies. One example is utilising physiological measurements in user testing, however, the usefulness of such approaches in improving games under-development has yet to be established. If applying physiological measure in GUR, one current challenge is tying together the results of physiological measures and player experience reports, because the data are different and identifying actionable results is difficult. Presenting results from high-resolution data to game designers is often not straightforward.

A common approach is to visualise large data sets captured directly as a result of gameplay (i.e., game metrics). These visualisations aim to analyse player performance aspects, such as player progress (e.g., time taken, location of death) or to balance gameplay (Medler, John, & Lane, 2011; Wallner & Kriglstein, 2012). However, most of these techniques focus on player behaviour (e.g., ‘what they did’) but do not address player experience, such as reasoning (e.g., ‘why they did it’), or emotion (e.g., ‘how they felt’). It is important to note that while analysis of player behaviour via telemetry cannot provide direct knowledge of experience, it is still possible to draw inference from behaviour to experience.

In summary, while classic UTs and physiological-based UT methods are increasingly being used in GUR, there are many open questions about both techniques and their value for improving game design. The study reported in this chapter aims to explore the use of two exemplar UT techniques in the development of a game, with each technique taking the game to

a different parallel version. First, a classic UT technique, for which the study combines gameplay observation, questionnaires and interviews as that they are widely used for creating a UT report. Second, a physiological-based UT technique (BioSt) that explores the use of storyboarding to visualise physiological data.

6.3 Overview of Evaluation

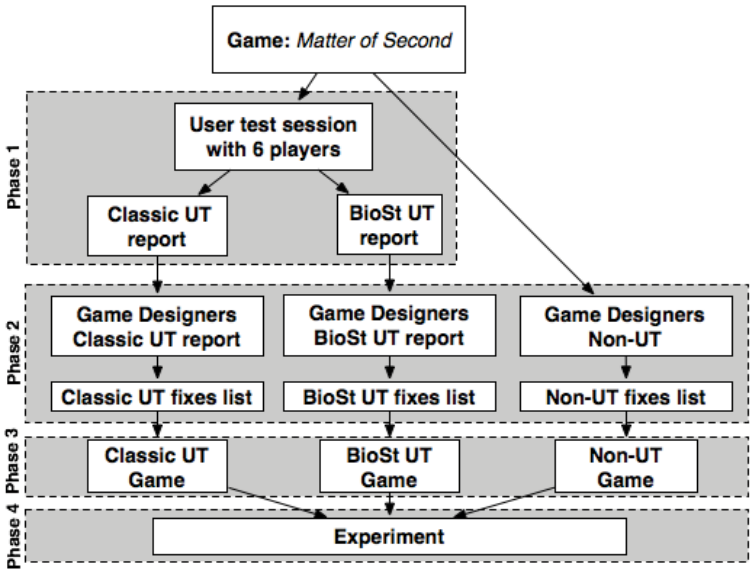


Figure 6-1 The overview of evaluation

The aim of the evaluation was to demonstrate whether different GUR techniques can facilitate the game design process to optimize the quality of the game. For this purpose, an under-development test game was selected from an indie game programmer. In *phase 1* (see Figure 6-1) a UT session was conducted on the test game using both a classic UT and a biometric (with BioSt) UT. In *phase 2*, three pairs of designers (see section 6.6: Phase 2 – Developing Three Game Prototypes for details) provided their improvement suggestions on the initial version of the game. One group of designers was given a classic UT report from Phase 1, the other group of designers used a BioSt UT report (from the BioSt tool explained in Chapter 5), a third (control) group of designers had to suggest design improvements without any UT reports based solely on their design intuition. In *phase 3*, the game programmer improved the game based on the design feedback. Each of these game prototypes was evaluated as a level of the independent variable in the experiment described in *phase 4*.

6.4 The Game: Matter of Second (MoS)

The game used in this is study was the independent game *Matter of Second*, a fast 2D platform jump-and-run game, under development (see Figure 6-2). The game programmer agreed to contribute to this project in return for receiving feedback to improve the quality of the game for its future release. In *MoS*, the main game goal of the hourglass-shaped player avatar is to stay

alive. The game is linear, as the player automatically moves forward to the game world with controls only allowing them to slow down, speed up, or jump over obstacles and gaps using three arrow keys. The player must collect items to add to the time left for completing a level while avoiding crashing. If the player dies they teleport back in time to the last checkpoint that they hit. If they run out of time, they must start over from the beginning of that level. Only the first two levels of the game were developed and completed to a playable quality indicative of the final release, and only these sections were tested, intended for about six minutes of gameplay.

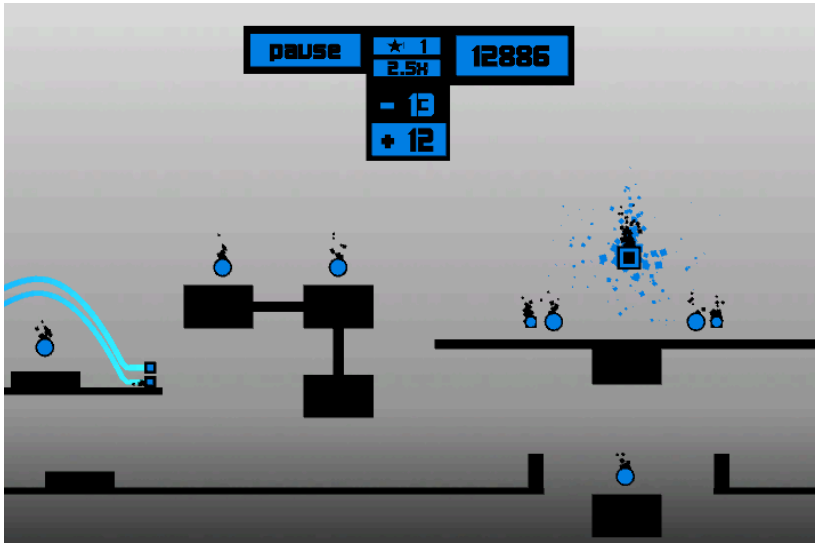


Figure 6-2 Screenshot of MoS level 1

6.5 Phase1 - User Test Sessions

For the initial user test sessions, six PC gamers were recruited, all male university undergraduate students, who played video games frequently. They were recruited using the department's internal mailing list, and their participation was voluntary. The aim of running the UTs was identifying usability and user experience (UX) issues with the initial game version.

Testing was conducted over 3 days in a game laboratory. Participants played the game on a PC (Windows 7) connected to a 24" display. Video cameras captured the player, physiological recording software recorded GSR and EMG signals, and real-time footage from the screen and simultaneous frame counter from the biometric software was digitally recorded for later analysis.

Before starting the UT, the players signed an informed consent form, and the GUR experimenter attached GSR and facial EMG sensors (see Figure 5-2 and Figure 5-3 in Chapter 5). The players were asked to relax for a few minutes so baseline measures could be recorded. The GUR experimenter informed the players that they would play the first two levels of a new game under development and their feedback was needed to improve the game. Each player played *MoS* for 10 minutes or two levels. Before the UT started the players were given a written

explanation of the game mechanics and had a few minutes to familiarise themselves with the controls. Once they felt comfortable, the GUR experimenter started the game, the video recording software and the physiological recorder. After they finished both levels (or 10 minutes of play) the recordings were stopped and the participant filled out a questionnaire asking them about their experience (e.g., game features they liked and did not like). After this, the players watched their gameplay video and were asked to pause the video to indicate and explain moments where they had any positive or negative gameplay experiences and provide qualitative comments on those moments. The GUR experimenter used the Player Input view of BioSt (as explained in Chapter 5) to enter the video timestamp and players' comments at the selected moments.

Two UT reports were then created by the GUR experimenter from the same UT session data: one, a classic text and video based UT report (this was later used by Classic UT team), and second, a BioSt UT report (this was later used by BioSt UT team).

The classic UT report included: the information about the game, the game's core mechanics, the participant's game profile details, the gameplay videos of the six participants, a check list of 9 gameplay issues identified from the user test sessions and a one-page description of each issue, followed by a screenshot, and examples from gameplay videos and participants' comment from the questionnaire.

The BioSt UT report included the same information about the game and participants as the classic UT report, the same gameplay videos of six participants, but instead of a description page of each issue and participants' comments from the questionnaire, it included a printed version of the BioSt GUR view, including the designer's intended player experience graph (see Figure 5-8 in Chapter 5), the player's GSR graph, the EMG positive and negative activations and the player's commented events (see Figure 5-9 in Chapter 5), and the aggregated player experiences graph (see Figure 5-10 in Chapter 5).

6.6 Phase 2 – Developing Three Game Prototypes

There were two reasons for conducting this development process. First, to create three parallel versions of the test game for the experiment (these versions served as 3 levels of the independent variable). Second, to evaluate the different approaches the game designers used for generating their recommendations list. Particularly to see the evaluation bandwidth (advantages and limitations) of each approach, how the designers' applied them, and to explore the difference in recommendations resulting from the different evaluation approaches.

For this six external game designers were recruited, all male graduates with a game development degree and with professional experience of developing video games. The designers

were selected carefully (same level of education and professional experience) to make sure they all had an equal level of expertise. They were also randomly grouped into three teams (two game designers per team) to reduce the effect of a designer's individual abilities. None of the recruited game designers were involved in the initial development of the game, neither were they aware of the study's aims nor of the BioSt tool. They were told that they were recruited to provide feedback on a game prototype for the programmer.

Each team was given the original version of *MoS*, the game's design intention, and a core game mechanics list that could not be modified (serving as publisher requirements). The designers were given 3 hours and asked to prepare their list of game improvement recommendations for the programmer. The designers were also asked to rate their confidence in each recommendation. Once they had prepared their list, they met the programmer explaining the requested changes. Each team (Classic UT, BioSt UT, and non-UT) attended their session on a separate day.

At the end of the session both game designers in the team were interviewed together. The interest was to know how and why they had come up with their recommendations list. Overall, designers working with BioSt requested the most changes (18, compared to 17 in Classic UT and 16 by Design Experts) and had the highest average confidence rating (4.8 out of 5 compared to 4.5 for Classic UT and 3.8 for Design Experts) among the groups.

All three groups modified the interface and the game's scoring system aiming to simplify them. In the Classic UT game and the non-UT game, the new scoring system was based on the number of collected items and the time a player stayed alive. However, in the BioSt UT game, the designers changed the scoring system to making players race against a timer so getting collectables reduced overall time.

The Classic UT team added a short tutorial at the beginning of level one, introducing the core game mechanics such as jumps. The BioSt UT team aimed to introduce these mechanics through the levels. For example, they asked to relocate the collectables showing a curve that players could use to make a good jump, or showing the ideal level path they should take.

Only the Classic UT team changed the high jumping mechanics and introduced a double jump. The BioSt UT team decided to modify the level, and omitted the platforms that needed a high jump. The BioSt game was the only game where designers changed the level design; adding to platform length before main challenges and after checkpoints.

The Non-UT team requested changes to give more feedback to players. For example, they asked for checkpoints to change colour once they had been activated.

The Classic UT and BioSt UT teams requested changes to the ending of each level. The Classic UT team added a static ending screen showing players overall time and score. The BioSt team requested a dynamic score bar which would activate based on the number of collected items. The players had to jump through the bar to reduce their overall time.

6.7 Phase 3 - Implementing Requested Changes

Once he had the three lists of suggested improvement changes, the game programmer then applied each set of changes to create three different parallel versions of MoS. He did not make any of his own modifications beyond these. This phase took about three weeks and the resulting three different versions of MoS were used as a condition for the experiment reported next.

6.8 Phase 4 - Experiment

In overview, to evaluate the relative value of the different GUR approaches, 24 participants (see below) played all three versions of the game: one developed using a classic UT text and video report, one developed by using a BioSt UT report, and one control condition developed only based on designers' opinions (see Table 6-1).

Conditions	***Development method***
Classic UT	Classic video and text UT evaluation and report
BioSt UT	BioSt UT evaluation and report
Non-UT	Designer's expert opinion

Table 6-1 Game conditions

6.8.1 Experimental Procedure

The study used a three-condition (2 UT variation approaches, 1 control without UT) within-subjects design. All participants played all three conditions, which were counterbalanced and presented using a randomized ordering. This means with the 24 participants in this study each counterbalanced sequence has been tested four times. After providing informed consent, the participants completed a demographics questionnaire, which also asked questions about their gameplay experience. Each participant was given a few minutes to get comfortable with the game controller (3 arrow keys) before the trial began. Participants played each game condition for 10 minutes or until they had completed both levels. Given the game context and the results form the initial user test sessions, 10 minutes is decided to be the right amount of time for the player to experience each game condition, this is also inline with a common playing time in other game research studies (Kivikangas et al., 2011a). After each game, participants completed four surveys (PANAS (Watson, Clark, & Tellegen, 1988), SAM (Bradley & Lang, 1994), SUS (Brooke, 1996) and a survey on the game features). PANAS questionnaire measures the two primary dimensions of mood- positive and negative affect. SAM is an easy to administer, non-

verbal questionnaire for quickly assessing the pleasure, arousal, and dominance associated a user's emotional reaction to an event. SUS is a simple, ten-item Likert scale questionnaire, providing a global view of subjective assessment of usability. Following completion of all conditions, players were interviewed and completed a final rating soliciting their opinions of the three games. The experiment was conducted over 7 days under laboratory conditions. The game was played on a Dell computer running Windows 7 with a 24" display. Participants were seated on a chair behind an office desk. They played the game using a standard Dell keyboard and speakers.

6.8.2 Participants

Twenty-four participants, all male students between 19 and 27 years old (M=23.3, SD=2.5) completed the study. They were all experienced PC gamers and played video games at least twice a month. Participants were recruited from a mailing list and received $10 for participating in the study.

6.9 Results

The ratings data were analysed using the Shapiro-Wilk test to check whether data were parametric or not. For the three related samples, significance was tested using a one-way repeated measures (RM) analyses of variance (ANOVA) for parametric data and using Friedman's ANOVA with an exact test for non-parametric data. This section presents quantitative results on the SUS, PANAS, and SAM scales, the preference rating scales, and results from qualitative interviews and observations of the game development process.

6.9.1 Results from SUS, PANAS, and SAM Scales

The data from the SUS was normally distributed, but the results were not significantly different between the conditions (F=2.83, p=.069). Positive affect (PA) and negative affect (NA) scores were calculated from the PANAS. For NA, the Friedman's test results were not significant. For PA, the data measured on a scale and show specific properties (Additively and linearity, normality of sampling distribution and residuals (error), homoscedasticity/homogeneity of variance, independence and interval or ratio level data), which allow us to use parametric statistical methods on the data and analysed with an RM ANOVA. The main results were significant (F=7.29, p=.002, effect size=0.241). Pairwise comparisons revealed significant contrasts between Classic UT and non-UT teams (mean difference=3.458, standard error=0.985, p=.006) and BioSt UT and non-UT teams (mean difference=4.375, standard error=1.211, p=.004). Figure 6-3 shows the differences between PA in the three different conditions.

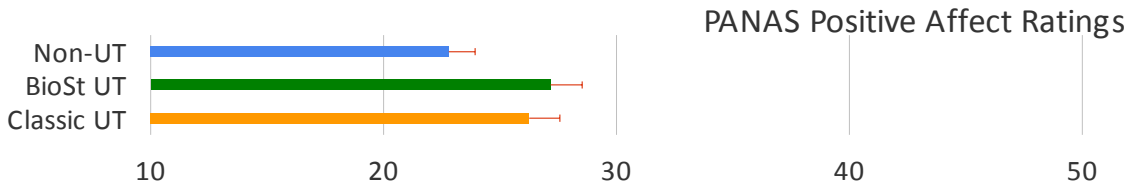


Figure 6-3 Significant mean (CI: 95%) PA rating from PANAS

SAM data was non-parametric and not significant for the arousal and dominance dimensions. However, for the SAM pleasure dimension, the Friedman's test results were significant (χ^2=12.2, p=.002). Pairwise Wilcoxon Signed-Rank tests showed that players found playing the Classic UT game (Z=-3.3, p<.001) and the BioSt UT game (Z=-2.2, p=.011) both more pleasurable than playing the Non-UT game, but no difference in SAM pleasure between the game versions created with Classic UT and BioSt UT (Z=-3.2, p=.386). Figure 6-4 shows the average values of the SAM scores and error bars at 95% confidence interval level.

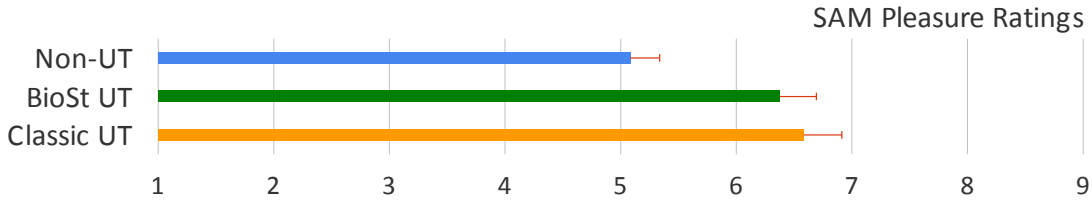


Figure 6-4 Significant average (CI: 95%) SAM Pleasure Rating

6.9.2 Results from Personal Preference Ratings

Participants also rated seven game attributes on a scale from 1 (worst) to 5 (best) for each level: Jumping, Time, Scoring, Controls, Speed, Collectables, and Level Design. All the data were non-parametric and a Friedman's test showed significant differences for time (χ^2=8.2, p=.015), scoring (χ^2=13.8, p=.001), and collectables (χ^2=13.1, p=.001). Pairwise Wilcoxon Signed-Rank tests showed that players again found no differences between playing the Classic UT and the BioSt UT games, but thought that both games improved *time* (both Z=-2.1, BioSt UT p=.012; Classic UT p=.024), *scoring* (both Z=-3.2, p<.001), and *collectables* (BioSt UT Z=-3.1, p=.001; Classic UT Z=-3.2, p<.001) compared to the Non-UT version.

At the end of the experiment after having played all 3 different game versions, participants were asked to make comparisons of the features in the three games (on a scale of 1 to 5). All of these features were significantly different in a Friedman's test (see Figure 6-5): Gameplay Experience (χ^2=11.4, p=.003), Gameplay Quality (χ^2=9.6, p=.006), Fun (χ^2=9.5, p=.008), Game Visuals (χ^2=7.0, p=.028), except for Game Sounds (χ^2=3.0, p=.667).

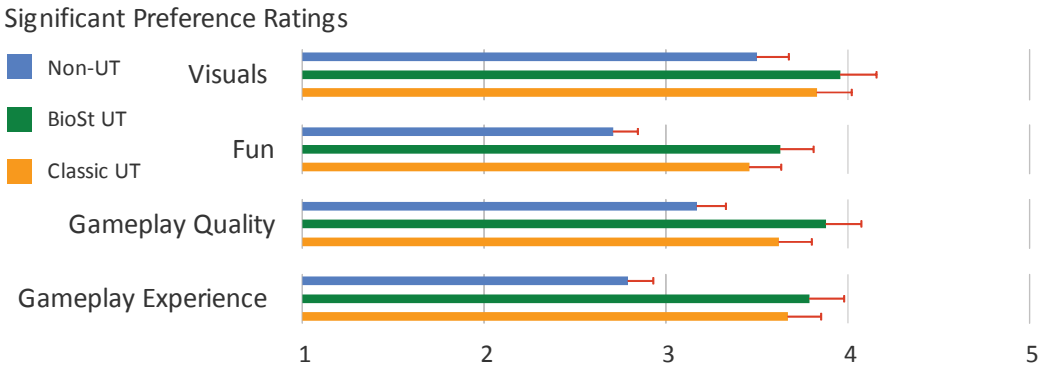


Figure 6-5 Significant average (CI: 95%) preference ratings

Again, this was followed up with Pairwise Wilcoxon Signed-Rank tests and for *gameplay quality*; no difference between Classic UT and BioSt UT games (Z=-1.2, p=.28), and no significant quality difference between games developed with Classic UT and Non-UT (Z=-2, p=.052) were found. But players rated the quality of the BioSt UT game significantly higher than the non-UT game (Z=-2.7, p=.005). The same results prevailed for *fun* – comparing BioSt UT to Classic UT (Z=-0.4, p=.73), comparing Classic UT to Non-UT (Z=-1.8, p=.08), and comparing BioSt UT to Non-UT (Z=-2.7, p=.008) – and *visuals* – comparing BioSt UT to Classic UT (Z=-1, p=.51), comparing Classic UT to Non-UT (Z=-1.6, p=.13), and comparing BioSt UT to Non-UT (Z=-2.6, p=.008). However, *gameplay experience* showed different results (both Classic UT and BioSt UT were not different from each other (Z=-4.9, p=.66), but both were significantly better than Non-UT (BioSt UT vs. Non-UT: Z=-2.9, p=.004); Classic UT vs. Non-UT: Z=-2.8, p=.005).

6.9.3 Results from the Players' Interviews

Players were interviewed after they had played all three versions of the game and they were able to compare them. The written notes from the open-ended interviews were analysed in over-reaching gameplay categories. This section provides some of the participants' comments from those interviews:

Jumps: Although most of the players mentioned they liked the idea of double jumping (recommended by the Classic UT team), they also criticized that the double jump was not well implemented and supported by the level design. For example, the game character (an hourglass) travels between two platforms in most levels without sufficient opportunities to perform a double jump. P3: "*I really liked the double jump, but level design did not recognize this [...]*." or P9: "*timing to do double jumps is not as intuitive as I expect*."

Speed: The players liked increasing the traveling speed of the hourglass, but the levels also required a player to slow down. An improved level design would use both slowing down and

speeding up more frequently as challenge mechanics in the game. For example P3:"*the controls and speed feels better in this version [non-UT Game], this is maybe because the camera zoomed in and zoomed out more when you go fast or slow, although the bad thing was when you slowed down the camera zoomed in too fast*."

Level design: The BioSt UT team was the only team that requested changes in the level design. For example, as mentioned earlier they added to the length of the platforms before the main challenges and after the checkpoints. These changes helped players toggling their speed before level challenges by giving them more time and space to make decisions. Players commented positively on the BioSt UT game's level design but did not explicitly notice these changes. For example, P5 mentioned: "*I don't know what is different in this version [BioSt UT game] but I just felt it had a better flow through the level. The level lets me to go fast, I could play the way I wanted to play*".

Game difficulty: Pace in levels was another important component picked up by players, for example P18 commented on the difficulty of the Classic UT game's levels after playing all three versions: "*the difficulty level was really high in some spots, if the difficulty was like [BioSt UT game] that would be perfect*." similarly P4*:"[BioSt UT game] levels get more difficult by progress.*"

Scoring and collectables: All three teams requested new implementations of the scoring system and integration of collectables. In Classic UT and Non-UT games, the scoring system was work-based. For example, based on the number of collected items and time a player stayed alive. However, in the BioSt UT game, the designers tried to simplify the scoring system by making players race against a timer so that getting collectable items reduced their overall time. Players found the scoring system in the BioSt UT game easier to understand. For example P7: "*In [Classic UT and Non-UT games] because the scoring system was just like points I didn't really pay attention to it, for me my primary goal was to finish the level. [BioSt UT Game] really gives you a goal to achieve trying to reduce the completion time.*" Players also commented that they felt a stronger motivation for getting more collectables in BioSt UT game. For example, it was observed that players were killing their avatar on purpose to be able to restart from previous checkpoints and get more collectable items.

Designers in the BioSt UT team used collectables for guiding players through a level by showing them the ideal curve to perform a jump or showing the correct path to platforms with more collectables. The interview transcripts showed that most of the players did not notice this guided use of collectables, but they commented that they thought the BioSt UT game had more collectables. They also felt they could perform better jumps. For example P5:"*in this version [BioSt UT] I knew how to jump, before I had no idea how far I could jump and how high I could*

jump, but in this one [BioSt UT] I had the parabolic motion, with all the power ups and minus time [collectables] so I knew when I should be able to jump to get them all and that really helped me at the beginning […]."

These – from a player's point of view – seemingly covert adjustments of the BioSt UT team had a great impact on a player's experience and affected their rating on the overall game quality.

6.9.4 Results from the Game Designers' Interviews

Each group of designers met the game programmer at the end of their session to explain their requested changes. These sessions were observed and then the game designers were interviewed after they explained their change requests to the game programmer.

Designers in the Classic UT team mentioned: "*we have fixed all the issues from the report.*" The researcher observed them referring to each issue when discussing their solutions (e.g., they said: "*Did we deal with issue 4?*").

The researcher observed designers in the BioSt UT team watching the gameplay video, pausing the video at the indicated moments and reading the player's comments. They referred to a specific player when discussing their changes, for example they said: "*this change is answering player 3's comment but what about player 4.*"

Although the BioSt UT report did not map the changes in a player's physiological measure to any specific emotion, the designers tended to interpret the issues referring to emotions for example: "*look this guy gets so frustrated here."*

The designers in the BioSt UT team were the only team that used their UT report while discussing their requested changes with the game programmer (e.g., to show him specifically where in the game they wanted to apply changes). The designers also referred to the physiological data to convince each other about the changes they wanted to make or to prove the existence of a problem.

In the interview with the designers in the Non-UT team, they mentioned that they had based their changes on game design concepts, for example they requested changes to give the player more feedback.

6.9.5 Results from the Game Programmer's Interview

The game programmer was also interviewed after he had applied all changes from all three groups. His comments showed that he felt that the designers in the BioSt UT team had a better understanding of players and the game. For example: "*I found changes from the Classic UT and Non-UT teams were be based more on personal opinions of the designers, the BioSt UT team really focused on the comments made by the players*." He added: "*The BioSt UT team had the*

most significant changes in their version; the changes were not necessarily more complicated but rather they simplify and polish existing mechanics. This may be because the data and feedback is only useful if you improve on existing functionality or totally remove it." He concluded by mentioning: "*Each group of designers came up with their own different take on the game. There were some similarities, for example, all groups found problems with the interface. Though I saw a difference in the BioSt UT team's changes being the least extreme, only removing small elements that players commented on being confusing. The Classic UT and Non-UT teams choose to make major changes to both of their UI's [user interface]*."

6.10 Discussion

Overall this study showed that the two different types of UTs optimised the design of the game compared to designing without UTs. The main findings are:

- While the usability of all three games was considered equal, the games designed with UTs evoked higher positive affect (PANAS) and pleasure (SAM) ratings from players.
- Gameplay experience was significantly better when using either Classic UT or BioSt UT to inform design recommendations.
- Fun, visuals, and gameplay quality were significantly better when using BioSt UT compared to using designers' intuitions only (Non-UT). These differences were not significant between Classic UT and Non-UT games.
- BioSt UT and Classic UT were not significantly different from each other in all players' ratings (see Figure 6-5). Both methods improved the game significantly.
- The findings from the interviews suggest that BioSts provided more focused feedback on the details of the level design mechanics that helped the designers focus on suggestions that would improve player experience.

Benefits of physiological measures: One of the main GUR challenges is to have a better understanding of player experience to identify issues with a game, especially where the players interact with a game in a way not intended by the game designers. Observing gameplay and interviewing players can provide a rich data source as part of a UT. Providing user researchers and designers with player's physiological responses and comments based on game events helps them have a better understanding of the player experience and provides additional evidence for the existence of a problem, which might not be uncovered by classic UTs, such as the level design changes suggested by the BioSt UT team in the study. Both (a better understanding and more evidences) can increase the researcher's confidence and the likelihood of an issue to be reported to and taken seriously by the game development team.

Gameplay experience is more than just usability: The usability of all three games was considered equal (based on the results from SUS questionnaire), but preferences and affect ratings differed significantly. This means even when the three versions of the game offer equal levels of usability, they were different in positive affect and pleasure. It can be argued that classic usability measures and scales do not provide enough information on player experience.

You need more than one round of tests: These results also suggest that a single round of user testing is not sufficient to lead to the best optimisation of a game. For example, the existence of the double jump feature in the Classic UT game appealed to most of the players, but they also commented that it was not well implemented or supported by the level design. Hence, an iterative process through various UT sessions will bring the game closer to designers' intentions.

GUR UTs provide different improvements to a game: GUR aims to let designers be creative while providing feedback on how players understand game design ideas. If this feedback shows that the idea is not well understood by the players, it may also provide information that allows the designer to improve the implementation of the original idea. This study showed how the two different UT approaches contribute to this in different ways. The BioSt UT supplied a naming and the location of the problems (e.g., "issue with jump") while the Classic UT included a page of explanation for each problem. It is possible to argue that the explanation in Classic UT could unintentionally prime the designer towards a specific solution. On the other hand, the BioSt can support the designer's creativity by visualising the problem and providing substantiation from the player's comments, the gameplay video and the changes in the player's physiological state, leaving designers to explore possible responses. Previous work on reporting GUR findings support the idea that designers want to explore where the gameplay problems are rather than being told a cookie-cutter solution. This also supports findings from interview studies with game developers to evaluate earlier prototypes of BioSt.

BioSt seems suitable for level design and difficulty: The BioSt team was the only team recommending changes in level design. This could suggest that providing designers with the information on a player's arousal level and comments in a level would give designers structural information on level design. This also seemed to work for pacing and influencing difficulty in the game used in this study, since pacing related directly to game difficulty. Structural game level information from BioSt seems to be useful for this.

GUR challenges: A challenge for GUR is that game designers can be resistant to trusting a UT report. As game designers usually spend many months developing their game, they can be defensive about acknowledging UX and GUR issues. This may be because the designers feel that the *GURs* are criticizing their design, since in a general UT report the inputs from *users* are often hidden in the text. It was observed that the Classic UT team wrote (in the report in front of

each reported issues) whether they agreed or disagreed with the issue, although this was not asked for by the researcher. This shows designers' tendency to believe or not believe the reported issues. On the other hand, the texts in BioSt reports are generated from players' comments. It was also observed the designers in the BioSt team were using the report to convince and argue for *their* recommended solution. For example, they were pointing to players' comments, or changes in a player's physiological state to support or convince their teammate to recommend a change.

Player-cantered design: Designers in the Classic UT team tended to use the report as a centre of their improvement process. It was noted that they frequently used phrases such as "*did we answer this issues*". In the interview at the end of their session, they also mentioned, "*we have answered all the reported issues.*" However, for the designers in the BioSt team, players were the centre of attention. It was noted that they frequently used phrases such as "*what about player X, would this change answer their comment*?" This would support the idea that BioSt provides a tool that keeps the players at the centre of design decisions, which is critical for acceptance of GUR reports.

Plausibility and persuasiveness are two important factors when reporting usability and UX issues (Lai-Chong Law, 2011). By visualising player's comments alongside the change in their physiological state, BioSt enables GUR and game designers to achieve these two important components of UT reports.

Limitations: The study presented in this chapter does have limitations. I recognise that having the game analysed and having recommendations made by designers who had no involvement in the game up until that time is not how UTs in game development normally work. However, this was necessary for the validity of the study so that all designers could provide suggestions that were equally weighted. Further, the game prototype, the small number of participants, designers individual abilities (although the study designed in order to reduce these) and short gameplay, do not allow us to conclude that a single UT approach leads to the creation of a better game or a better experience for the player, although the findings indicate that BioSt has clear potential for improving fun, visuals and gameplay quality. Blending Classic UTs and BioSt UTs further could be an ideal solution for further research.

However, this study provides the opportunity to observe the evaluation bandwidth for each UT approach. Hence, it shows how designers perform when applying these different methods to make their design decisions (using them towards plausibility and persuasiveness).

6.11 Conclusion

This chapter addresses techniques that have already been applied in the GUR field, reassuring their value is relevant for this growing discipline. But the most important contribution is how those methods can inform and change the design practice.

The study presented in this chapter supports employing UTs in game development since they will yield a higher quality game and a better gameplay experience. A successful game will keep the player engaged and succeed at its intended purpose, whether it is to entertain or to inform. Hence, the findings not only apply to the game industry for improving entertaining games, but also to research seeking to create an enjoyable interactive system. Using either classic UTs or BioSt UTs (in combination with classic UT) will improve the user experience of these applications. Furthermore, this study provides initial evidence that BioSts provide more nuanced design feedback and provoke more subtle changes to game mechanics, which will result in higher perceived gameplay quality and essentially more fun in gaming applications.

6.12 Summary

This chapter demonstrates through a comparative study, how BioSt can assist GUR to gain a better understanding of player in-game behaviour and how this approach can facilitate communication of player experience issue to the game development team. The study also provides evidence as to how user test techniques in general can lead to a better gameplay experience.

Chapter 7 presents the thesis conclusion, revisits and discusses the research questions as well as offers some suggestions for future research in the GUR field. Chapter 7 also provides an overall thesis discussion.

7 Discussion, Conclusion and Future work

7.1 Summary

This thesis showcased the development of the Biometric Storyboards methodology through a series of studies that investigated the contributions of physiological measurements (GSR and facial EMG) together with qualitative evaluations in GUR. Chapter 2 demonstrated that, due to the specific characteristics of video games (see Table 2-1) classic user research methods need to be adjusted for GUR. For example, the need of continuous and unconscious measurements in user testing to better capture, analyse and report on players' experience, which builds the necessary foundation for this research. Building on prior research on the understanding of physiological measurement as a method for game evaluations, the studies conducted as part of this thesis expand the application of physiological game evaluation in GUR to provide iterative and formative feedback for game developers on games in all stages of development. The development of BioSt as an industry-focused methodology and tool that leverages most recent research in physiological game evaluation is documented throughout this thesis.

This final chapter of this thesis provides the summaries, discussions and contributions of each study, prototype evaluation, tool development and the final experiment. This chapter also revisits research questions, considering the contributions across the thesis. Finally, future work in this field is suggested.

7.1.1 Study One

This study discussed the value of physiological GUR approaches as an addition to traditional behavioural observation methods. The respective contributions to producing formative feedback during the development of video games were outlined. The results showed that observation-based techniques can expose the majority of usability issues. However, the biometrics approach enabled researchers to discover latent issues in players' feelings, their immersion and gameplay experience. Biometrics can also give researchers confidence, confirmation, and validation of issues.

7.1.2 Study Two

This study gave insight into player interaction and motivation by furthering previous social interaction research, combined with evaluation of player's physiological measurements. This study showcased a methodology for recording biometric responses and social interaction data to

gain greater insight into the motivations of players during collocated gaming sessions. The mixed methods were then used to study 16 players, across 8 sessions and consisted of the triangulation of physiological measurements, social interaction coding, self-assessment diagrams and player interview to understand how forms of social interaction resonate with specific player types. The results of the study advanced an understanding of the motivations behind player's interactions during collocated gaming; this advances previous work on social interaction in multiplayer gaming as well as demonstrating how physiological measurements can be successfully applied in combination with other user research approaches to better understand player experience.

7.1.3 Case Studies: BioSt Prototypes Iteration

The development of BioSt prototypes was reported in Chapter 4, where three prototypes were created iteratively based on user testing reports for two commercial console games that were in development. These prototypes explored how game developers used the BioSt visualisation report and what can be done to provide a sufficient and actionable data for game developers to create better gameplay experiences in the final release of the game.

7.1.4 Study Three: BioSt Prototypes Evaluation

This study evaluated the usefulness of BioSt to the game industry. The three prototypes of BioSt were presented to six professional game developers and they were interviewed about the advantages and disadvantages of this technique. The results suggest that BioSt can be used as a tool to enable discussion. They are easily understandable and use neutral language, resulting a tool that allows game developers to quickly pinpoint areas of the game that are working as intended and those that need to be refined. The interview results also highlighted areas of further development for the tool, such as generating a composite graph, including new sensors to measure different responses, and providing a comparison between game designers intended experience graph and players' BioSt graph.

7.1.5 BioSt Tool

The purpose of the BioSt tool is to visualise the data gathered from user test sessions including GSR, facial EMG measurements as well as players' comments. The tool enables researchers to compare the players' data with game designers' intended player experience, and also to combine these datasets into a single view that can later be shown to game designers. Using the tool, GURs can create an aggregated graph representing GUR findings from a number of players.

7.1.6 Study Four: BioSt Evaluation

This study demonstrated how standard user testing reports and BioSt both help designers create a better gameplay experience. The study showed that employing user testing in game development will yield a higher quality game and a better gameplay experience, and that BioSt

provide more specific and actionable design feedback that encourages more refined changes to game mechanics. The design implication is that a game designed with the BioSt method will result in higher gameplay quality. These findings can also apply to the broader entertainment industry, whereby using BioSt in user testing will improve the quality of entertainment systems by bringing them closer to designers' intentions.

7.2 Thesis Discussion

Each study presented in this thesis includes discussion respective to each study's content and conclusion, wherein each chapter reflects upon the process of the study, the challenges of formative rather than summative evolution, argued for the study's individual contributions, and positioned each study's research objectives. Such as, Section 3.2.4 which discussed the methodology, quality and quantity of gameplay issues and the use of GSR in combination with post-session interviews in study S1; Section 3.3.7, which provides relevant discussion for study S2, including the contribution of the approach to game developers, and game development; Section 4.7, which discussed the results from BioSt prototype evaluation, the importance of following UCD stages for the prototypes development, and how the design of BioSt iterated based on game developers' needs. Finally, Section 6.10 provided discussion based on study S4 (the main study of the thesis) which explores the benefits of physiological measurements in applied GUR studies and challenges for conducting GUR. In combination, these studies and bridging discussion sections form the discussion content for this thesis, and are concluded by additional discussion in this section (7.2).

One of the challenges for GUR is to have a better understanding of player experience to identify issues with a game. Observing gameplay and interviewing players can provide a rich source of data but they can be time consuming and indicating specific and actionable issues from these qualitative sources can be difficult. This thesis showcases one of the contributions of physiological measurements in GUR on pointing out significant moments in gameplay, which allows observation and post-session interviews to be made more efficient by only focusing on those selected moments. This thesis provides evidence that physiological measurements can provide a structure for observations and interviews based on game events with the greatest impact on players' feelings.

Utilising physiological measurements in triangulation with other use research methods provide extra source of data to identify (or confirm the existence of) usability and user experience issues. Such measures also increase the researcher's confidence and the likelihood of an issue to be reported to and taken seriously by the game development team. Plausibility and persuasiveness are two important factors when reporting usability and UX issues. BioSt is a step towards answering the need for a fast and simple analysis and representation model of physiological

measurements to communicate ideas on player experience. By visualising player's comments alongside the change in their physiological state, BioSt provides a meaningful representation of physiological data and components of player experience for GURs and game designers to achieve these two important components of user test reports.

Moreover, indicating positive game events allows developers to better understand successful elements of their game, and can be considered to be as useful as finding negative issues. This also suggests that, even when the identification of usability issues would be possible by classic user testing approaches, it will be difficult to identify issues related to positive affect and pleasure. It can be argued that classic usability measures and scales do not provide enough information on player experience.

Another challenge for GURs is that game designers can be resistant to believing in user test findings. Game designers usually spend considerably more time with their game in comparison to GURs, this can potentially lead to game designers self-justification instead of acknowledgment of UX and GUR issues. This challenge can also be greater if designers feel that the *GURs* are criticising their design, especially in a general user test report, the inputs from individual *users* are often hidden in the text. In order to address this, BioSt emphasises on players' input by visualising the change in their physiological measurements alongside with their comments.

Visualising player experience makes difficult-to-interpret GUR data more accessible to a wider game industry audience. BioSt utilises a correlation between user research data and gameplay events, and provides a visualisation for tying the GUR findings together, since these findings contain data with different formats (such as qualitative user comments, quantitative game metrics). BioSt represents how specific game mechanics resonate with player experience and behaviour; understanding how player experience changes by performing particular tasks in gameplay environments is vital information for game designers.

GURs have the benefit of understanding the games development process and the relevant needs in the working environment in order to design visualisations which closely match the requirements and language of target users, and the subsequent level of detail necessary for the task. BioSt enable increased collaboration between GURs, games designers, games developers and producers. The game designers and programmers contributed to the studies S3 and S4 commented that they were able to more effectively discuss design strategy using these storyboards as evidence for player behaviour.

As discussed in Chapter 2 (Table 2-2), Fulton, Ambinder, & Hopson (2012) have defined an evaluation framework and criteria to conduct applied GUR studies with a focus on formative

evaluation. In Table 7-1 I have revisited these criteria with the aim of evaluating the BioSt visualisation tool using this framework.

Representative	• Representative method should be applied • Representative participants should be recruited
Accurate	• BioSt provides and uses multiple sources of data • Uses unconscious physiological measurements • Precise equipment for physiological measurements • Continuous and quantitative data capture (not based on self-reports)
Specific	• BioSt has been designed to pinpoint gameplay issues
Timely	• BioSt tool facilitates analysis of physiological measurements and creation of BioSt in a more timely fashion • BioSt shares many similar aspects of writing user test reports • Suitable add-on to text/video reports
Cost-effective	• BioSt tool facilitates analysis and interpretation of physiological data • Comparable in length and cost of GURs training to use other methods such as interviews or observations • Commercial physiological measurement equipment are becoming more affordable • Physiological data can often be collocated in conjunction with other methods (e.g. Observation, which was presented in S1)
Actionable	• BioSt displays the location and scope of each identified issue • BioSt helps game designers to priorities issues to tackle
Motivational	• BioSt enables GURs and game designers to achieve a better understanding of player experience • BioSt shows the effect of issues on players' physiological measures (body) • BioSt provides extra evidence for existence of gameplay issues

Table 7-1 Evaluating BioSt based on criteria for applied GUR studies

Accurate: Physiological measurements provide precise data for BioSt. Moreover, BioSt delivers and uses multiple sources of data. Both (multiple sources and precise data) increase the accuracy of results.

Specific: Based on the results from the case studies and S3 and S4, BioSt has been designed to pinpoint gameplay issues; and this is on of the strength of this approach.

Timely: One of the aims of the BioSt tool is to facilitate creating BioSt in a more timely fashion. The tool has significantly reduced the BioSt creation time (from 1-2 graphs per day in case study prototypes to 6 graphs per day for the study S4). Also many aspects of generating BioSt are similar to writing user test reports and it would be a suitable add-on to text/video reports.

Cost-effective: The BioSt tool has been designed to facilitate analysis and interpretation of physiological data for GURs, training personnel to use physiological evaluation would have similar length and cost of training them to conduct user test sessions and analysis of for example observation or interview data. Moreover, commercial physiological measurement tools are becoming popular and affordable.

Actionable: By displaying the location and scope of each identified issue, BioSt helps game designers to priorities and tackle the issues. The results of the case studies also indicated that these are the key data points for game designers to be able to act on the issues.

Motivational: By visualising player's comments alongside the change in their physiological state, BioSt enables GUR and game designers to achieve a better understanding and have more evidence for an existence of an issue. These can increase the user researcher's confidence and the likelihood of the issue to be reported to and fixed by the game development team.

7.3 Thesis Contributions

The goal of this research is to introduce a quick, and easy-to-understand method that integrates physiological data into GUR. The main contributions of this research can be summarised in three areas; firstly, incremental improvement of classic user research techniques (such as self-report, storyboards, physiological analysis); secondly, visualising player experience through changes in physiological signals; and finally, deconstructing game design by analysing game pace and events. The following section outlines each specific contribution:

7.3.1 Using Player's Physiological Measures to Structure Post Gameplay Interview

A novel combination of physiological measures with existing user research approaches is introduced in study S1. In this approach, changes in player's GSR were utilised to timestamp micro-events from their gameplay with greater effect on player's feelings. These selected micro-events were then used to give structure to the post-session interview with the player after their gameplay session ended. Although previous research has used physiological evaluation as a measure for dependent variables, there has been no previous research investigating the application of physiological measurements to identify potential individual usability and user experience issues in game development cycles.

7.3.2 Deconstructing Game Design by Analysing Pace and Events

Selecting events based on their effect on player's feeling brought a new prospective for deconstructing game design. In addition to identifying player behaviour from the gameplay video, the use of physiological measurements also shows the corresponding physiological reaction from the player's body. Thus, the technique presented in 7.3.1 forms a structure for analysing the coupling between player behaviour and feeling.

7.3.3 Using Player's Physiological Measures to Visualise Their Gameplay Experience

Various visualisation techniques aim to provide a better understanding of player experience, however most of these techniques are based on players' behaviour or actions, and do not include players feelings resulting from those actions. By using a mixture of qualitative and quantitative approaches, BioSt is designed to use both groups to motivate scientific discussion and bring the field forward.

7.3.4 BioSt Tool and Method for Analysing Player Experience

This thesis focuses on the BioSt as a novel approach, which is still evolving and being refined, but has shown potential in providing adequate data and enhancing understanding of player experience for game developers to effectively improve their game. Chapter 4 addresses how BioSt has improved as a methodology. This work builds from employing the techniques and dives deeply into how they evaluate the user differently, and how designers utilise that understanding. The aim is to introduce a method that contributes to a deeper understanding of player experience.

7.3.5 Guidelines for Reporting GUR Findings

Results from iterative prototyping and the interview study S3 with professional game developers to evaluate BioSt, studied important information on game developers' expectations from UT's report. The study identified and discussed 8 important aspects of GUR reporting that can provide guidelines for further studies in the field: (1) At a glance summary (2) Objective credibility (3) Location and prioritising of gameplay issues (4) Identifying a problem/suggesting a solution (5) Clarity/simplicity (6) Facilitates the discussion (7) Trust/convincing (8) Comparison to intended experience.

7.3.6 Systematic Explanation on How GUR Can Help Improve Gameplay Experience

The methods presented in this thesis (simple iterative design, classic user testing, and physiological measurements) have already been used in the GUR studies, however this thesis brings a relevant contribution by experimenting with how these techniques affect the understanding of the user as well as the GUR process itself.

The studies reported in this thesis are designed to look at the whole GUR cycle (from user test session, to UT report, to game) to discuss the evaluation bandwidth of each approach. As such, study S4 provided an experimental conditions to observe and evaluate how game designers performed when applying findings from different GUR approaches, how they made their design decisions as well as how game designers communicated GUR findings to game programmers, a critical stage of the GUR cycle.

7.4 Limitations and Future Work

7.4.1 Including Game Analytics Data

Dividing the game into segments based on gameplay and visualising the physiological data and player comments for each brings a new perspective to user research data, which was not available for GURs and developers before. BioSt can be part of a framework to include other gameplay data within a similar setting. For example, it is possible to look at each player avatar's death locations in a given segment of gameplay. Together with physiological data and player comments (see Figure 7-1 for an example prototype). One possible further develop of BioSt tool could be to include the mapping between game metrics and physiological data and player comments.

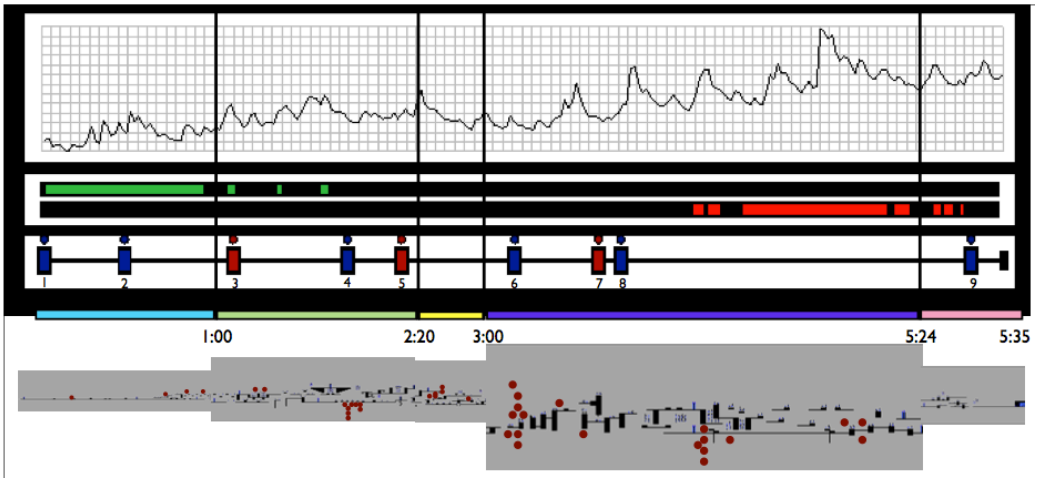


Figure 7-1 Prototype idea to include map of player avatar's death locations (red dots) into BioSt

Another possible development would be experimenting with different visualisations with the same goal to triangulate multiple data sources, for example: movement data, players' verbal comments from interviews, and physiological data. In particular, the combination of physiological data with spatial and temporal markers provided by the game (e.g., mapping physiological measurements to a player's path to provide us with a better understanding of the measurements in relation to the in-game environment) would be a potential next step building on the results from this thesis.

7.4.2 Improvement on BioSt Tool

Further research into using differing physiological sensors may suggest that specific sensors, or sensors used in combination, can reveal yet greater number of issues. The current BioSt tool uses GSR to measure player's arousal and facial EMG as measure of valence. Further improvement into inclusion of differing physiological sensors, such as EEG for looking at relax or focus states, may allow more of the player emotion spectrum to be represented. Another potential improvement (current limitation) would be creating an automated system that allows peak detection and reporting of phasic physiological responses at game events.

7.4.3 Further Study into Contributions of Biometrics in GUR

Blending classic user testing and physiological measurements further could be a possible next step for a future study to reveal constructive differences of each approach, building on the results from the work presented in this thesis. A more qualitative study is essential to better reveal the constructive differences of each approach. This reflects on some of the on-going work that I am already conducting as part of the next steps. For example, a follow-up study with multiple designers on this issue is worth suggesting for a potential future work in this domain.

7.4.4 Framework for Summative Evaluation of Player Experience

The work presented in this thesis shows the value of obtaining and representing players' physiological measurements alongside with their interview comments and game events, to provide a formative report in order to improve games under development. However, the approach presented in this thesis can be iterated to provide a framework for summative evaluation of player experience. For example, to create BioSt to compare gameplay journeys of different players and use them to spot key trends (players' internal factors that motivate their gameplay) in order to better understand their behaviour and experiences.

The synchronised representation introduced in this thesis deconstructs the gameplay session alongside players' physiological measurements. One of my further interests is to apply this detailed analysis of gameplay sessions to identify potential relationships between different elements of gameplay and their effect on long-term (resonated) player experience.

7.4.5 Other Applications:

The proposed approach and findings not only apply to the GURs and game industry professionals for improving entertaining games, but also to people seeking to create games around simulation environments, evaluating user experiences in entertainment technologies (such as movies and advertisement domains), or wanting to "*gamify*" applications.

7.5 Conclusion

While the game industry is facing fast changes in the market as well as advances in game technology and rise in game development costs, studios are under pressure to ensure their games

are successful. One approach, which has increasingly been applied to make sure players experience the game as the designers' intended is that of GUR.

Based on requirements from the game industry, GUR methods have been adapted and evolved from HCI evaluation methods to provide a mixture of qualitative and quantitative approaches for evaluators to choose from depending on their goal (see Figure 7-2). However, identifying the effective mixture of these methods and mapping (blending) the results from each of them together into a meaningful, actionable and easy to understand report to be presented to game designers, is one of the current challenges facing GUR.

Figure 7-2 Current methods of GUR

The BioSt method presented in this thesis has been developed using a mixture of qualitative/quantitative and subjective/objective methods to provide a powerful evaluation tool of player experience (Figure 7-3).

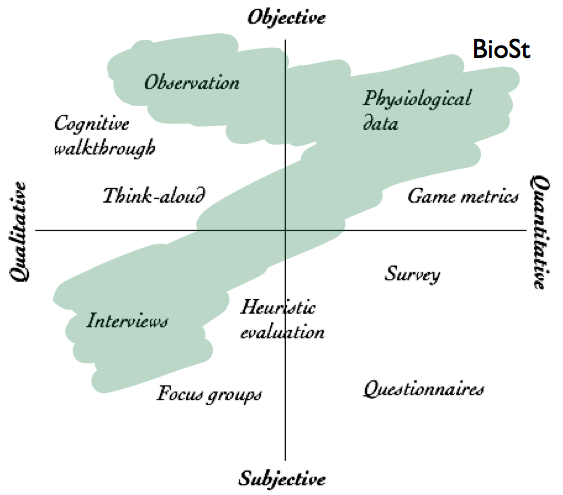


Figure 7-3 BioSt combines objective/subjective and qualitative/quantitative evaluation methods in a single mixed method

One of the key advantages of this approach is that the output is visual. This provides suitable feedback for the developers as they can quickly scan for key differences in level design, player performance and change in player emotions. Providing easy to interpret feedback to developers is a strong advantage of this user research method.

Although BioSt uses physiological measurements as an underlying technique to identify gameplay issues, understand player motivations, and report the player experience, they also have limitation (as discussed earlier they are highly subjective measures prone on several uncontrolled factors like caffeine, noisy signals, habituation, etc.). No single user research method is, for that matter. The difference is that physiological measurements provide another data point for BioSt with which to better inform an opinion, to help GURs and game developers to reduce the amount of uncertainty in explaining an issue and, to add confidence to the findings in user test reports. BioSt provides reasonable set of data points to make an informed decision on the player experience.

Another contribution of this thesis resides in assessing the impact of different user research methodologies to video game development. Considering that GUR continues adjusting HCI techniques into the entertainment sphere, this work moves forward the discussion by addressing how the selected techniques inform designers and consequently make a different impact to the design process.

To summarise, video games are highly complex, but still not as complex as understanding the human (player) behaviour. User research methods need to evolve, and this thesis tackled the problem of improving qualitative evaluation of player experience within GUR. Biometric Storyboards have the potential to change the face of games user research because they are quick and easy-to-understand, but also visualise complex user research data to produce an accurate, specific, actionable and motivational representation of user test findings suitable for the games user research process.

Bibliography

Albert, W., & Tullis, T. (2013). *Measuring the User Experience*. Morgan Kaufmann.

Ambinder, M. (2011). Biofeedback in Gameplay: How Valve Measures Physiology to Enhance Gaming Experience. *Game Developers Conference 2011*.

Andersen, E., Liu, Y.-E., Apter, E., Boucher-Genesse, F., & Popović, Z. (2010). Gameplay analysis through state projection. Presented at the FDG '10: Proceedings of the Fifth International Conference on the Foundations of Digital Games, ACM. doi:10.1145/1822348.1822349

Andreasen, E. S., & Downey, B. A. (2003). Measuring Bartle-quotient. *andreasenorg*. Retrieved April 13, 2011, from http://www.andreasen.org/mud.shtml

Axelrod, L., Fitzpatrick, G., Henwood, F., Thackray, L., Simpson, B., Nicholson, A., et al. (2011). "Acted reality" in electronic patient record research: a bridge between laboratory and ethnographic studies (Vol. Part II , Volume Part II). Presented at the INTERACT'11: Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction, Springer-Verlag.

Balaam, M., Fitzpatrick, G., Good, J., & Harris, E. (2011). Enhancing interactional synchrony with an ambient display. Presented at the CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. doi:10.1145/1978942.1979070

Barendregt, W., & Bekker, M. M. (2006). Developing a coding scheme for detecting usability and fun problems in computer games for young children. *Behavior research methods*, *38*(3), 382–389. doi:10.3758/BF03192791

Bartle, R. A. (1996). Richard A. Bartle: Players Who Suit MUDs. *Journal of MUD research*.

Bernhaupt, R. (Ed.). (2010). *Evaluating User Experience in Games: Concepts and Methods (Human-Computer Interaction Series)* (2010 ed.). Springer.

Blythe, M. A., Overbeeke, K., Monk, A. F., & Wright, P. C. (2004). *Funology*. Kluwer Academic Pub.

Boucsein, W. (1992). *Electrodermal Activity*.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59. doi:10.1016/0005-7916(94)90063-9

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624–634. doi:10.1016/j.jesp.2009.02.016

Bromley, S. (2012). Capturing Fun: creating a tool to measure social interaction during play testing. *CHI '12: Proceedings of the Game User Research Workshop*, 1–4. Retrieved from http://hcigames.businessandit.uoit.ca/chigur/wp-content/uploads/2012/04/gurchi2012_submission_35.pdf

Bromley, S. (2011, September 18). *Beyond Trash Talk: Understanding player motivation through analysis of social interaction in collocated multiplayer gaming*. University of Sussex 2011.

Bromley, S., Mirza-Babaei, P., McAllister, G., & Napier, J. (2013). Playing to Win? Measuring the Correlation Between Biometric Responses and Social Interaction in Co-located Social Gaming. In Multiplayer: *The Social Aspect of Digital Gaming* (T. Quandt and S. Kroeger, Eds.) Routledge. ISBN 978-0-415-82886-4.

Brooke, J. (1996). *SUS-A quick and dirty usability scale*. Usability evaluation in industry.

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. *CHI EA "04: CHI "04 Extended Abstracts on Human Factors in Computing Systems*. doi:10.1145/985921.986048

Bødker, S. (2006). When second wave HCI meets third wave challenges. Presented at the NordiCHI '06: Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, ACM Request Permissions. doi:10.1145/1182475.1182476

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). *Handbook of Psychophysiology* (3rd ed.). Cambridge University Press.

Canossa, A., & Cheong, Y. G. (2011). Between Intention and Improvisation: Limits of Gameplay Metrics Analysis and Phenomenological Debugging. *Proceedings of DiGRA*.

Chanel, G., Rebetez, C., Bétrancourt, M., & Pun, T. (2008). Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. Presented at the MindTrek '08: Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era, ACM. doi:10.1145/1457199.1457203

Church, D. (1999). Formal abstract design tools. *Gamasutra Features*. Retrieved from http://www.gamasutra.com/view/feature/131764/formal_abstract_design_tools.php?

Costikyan, G. (2002). I Have No Words & I Must Design: Toward a Critical Vocabulary for Games. *CGDC Conf.*

Crothers, B. (2011, October 14). Storyboarding & UX – part 1: an introduction | Johnny Holland. *johnnyholland.org*. Retrieved June 13, 2012, from http://johnnyholland.org/2011/10/storyboarding-ux-part-1-an-introduction/

Desurvire, H., & Wiberg, C. (2009). Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games: The Next Iteration. Presented at the OCSC '09: Proceedings of the 3d International Conference on Online Communities and Social Computing: Held as Part of HCI International 2009, Springer-Verlag.

Dixit, P. N., & Youngblood, G. M. (2008). Understanding playtest data through visual data mining in interactive 3d environments (pp. 34–42). Presented at the CGAMES 2008.

Drachen, A., & Canossa, A. (2009a). Analyzing spatial user behavior in computer games using geographic information systems. Presented at the MindTrek '09: Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era, ACM. doi:10.1145/1621841.1621875

Drachen, A., & Canossa, A. (2009b). Towards gameplay analysis via gameplay metrics. Presented at the MindTrek '09: Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era, ACM Request Permissions. doi:10.1145/1621841.1621878

Drachen, A., Canossa, A., & Sørensen, J. R. M. (2013). Gameplay Metrics in Game User Research: Examples from the Trenches (pp. 285–319). In Seif El-Nasr, M., Drachen, A., & Canossa, A. (Ed.), *Game Analytics: Maximizing the Value of Player Data*. London: Springer London. doi:10.1007/978-1-4471-4769-5_14

Drachen, A., & Smith, J. H. (2008). Player talk—the functions of communication in multplayer role-playing games. *Computers in Entertainment (CIE*, *6*(4). doi:10.1145/1461999.1462008

Drachen, A., Nacke, L. E., Yannakakis, G. N., & Pedersen, A. L. (2010a). Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. Presented at the Sandbox '10: Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games, ACM. doi:10.1145/1836135.1836143

Drachen, A., Nacke, L. E., Yannakakis, G. N., & Pedersen, A. L. (2010b). Psychophysiological Correlations with Gameplay Experience Dimensions. Presented at the CHI 2010 Workshop: Brain, Body, and Bytes.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*.

Fabricatore, C., Nussbaum, M., & Rosas, R. (2002). Playability in Action Videogames: A Qualitative Design Model. *Human-Computer Interaction*, *17*(4), 311–368. doi:10.1207/S15327051HCI1704_1

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting with Computers*, *21*(1-2), 133–145. doi:10.1016/j.intcom.2008.10.011

Fairclough, S. H. (2011, July 27). Biometrics and evaluation of gaming experience part two: a thought experiment. *Physiological Computing where brain and body drive technology*. Retrieved July 9, 2013, from http://www.physiologicalcomputing.net/?p=1760

Federoff, M. A. (2002). Heuristics and usability guidelines for the creation and evaluation of fun in video games.

Feigenbaum, E. A., & Simon, H. A. (1962). A THEORY OF THE SERIAL POSITION EFFECT - FEIGENBAUM - 2011 - British Journal of Psychology - Wiley Online Library. *British Journal of Psychology*.

Fowles, D. C. (1986). The eccrine system and electrodermal activity. In M. G. H. Coles, E. Donchin, & S. W. Porges (Eds.), *Psychophysiology* (pp. 51–96). New York: Guilford Press.

Frauenberger, C., Good, J., Alcorn, A., & Pain, H. (2012). Supporting the design contributions of children with autism spectrum conditions. Presented at the IDC '12: Proceedings of the 11th International Conference on Interaction Design and Children, ACM. doi:10.1145/2307096.2307112

Free Radical Design. (2008). *Haze*. Ubisoft 2008.

Fridlund, A. J., & Cacioppo, J. T. (1986). Guidelines for human electromyographic research. *Psychophysiology*.

Fullerton, T. (2008). *Game Design Workshop*. Taylor & Francis US.

Fulton, B., Ambinder, M., & Hopson, J. (2012). Beyond Thunderdome: Debating the effectiveness of different user-research techniques. In B. Fulton (Ed.). Presented at the IGDA GUR SIG Summit 2012. Retrieved from http://vimeo.com/groups/gursig/videos/26733185

Gingras, J. (2012). The POWER of VISUAL CONSENSUS in a MULTIDISCIPLINARY ENVIRONMENT. Presented at the MIGS 2012, Montreal.

Glaser, B. G., & Strauss, A. L. (2009). The discovery of grounded theory: Strategies for qualitative research.

Good, J., Howland, K., & Nicholson, K. (2010). Young People's Descriptions of Computational Rules in Role-Playing Games: An Empirical Study (pp. 67–74). Presented at the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), IEEE. doi:10.1109/VLHCC.2010.18

Gow, J., Cairns, P., Colton, S., Miller, P., & Baumgarten, R. (2010). Capturing Player Experience with Post-Game Commentaries. Presented at the CGAT Conference 2010.

Hakner, J. (2009, December 3). A Vertical Slice of the action. *A Vertical Slice of the action*. Retrieved May 21, 2013, from http://www.sussex.ac.uk/staff/newsandevents/newsarchive?id=2515

Hassenzahl, M. (2005). The Thing and I: Understanding the Relationship Between User and Product. In *Funology: from usability to enjoyment* (Vol. 3, pp. 31–42). Dordrecht: Kluwer Academic Publishers. doi:10.1007/1-4020-2967-5_4

Hazlett, R. L. (2008). Using Biometric Measurement to Help Develop Emotionally Compelling Games. In K. Isbister & N. Schaffer (Eds.), *Game Usability: Advancing the Player Experience*. Morgan Kaufmann.

Hazlett, R. L., & Benedek, J. (2007). Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, *65*(4).

Hoobler, N., Humphreys, G., & Agrawala, M. (2004). Visualizing Competitive Behaviors in Multi-User Virtual Environments. Presented at the VIS “04: Proceedings of the conference on Visualization ”04, IEEE Computer Society.

Hot, P., Naveteur, J., Leconte, P., & Sequeira, H. (1999). Diurnal variations of tonic electrodermal activity. *International Journal of Psychophysiology*, *33*(3), 223–230. doi:10.1016/S0167-8760(99)00060-4

Hullett, K., Nagappan, N., Schuh, E., & Hopson, J. (2011). Data analytics for game development (NIER track). Presented at the ICSE '11: Proceeding of the 33rd International Conference on Software Engineering, ACM. doi:10.1145/1985793.1985952

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. Presented at the Challenges in Game AI Workshop, 19th National Conference on Artificial Intelligence, San Jose.

Ijsselsteijn, W. A., de Kort, Y., & Poels, K. (2008). Toward real-time behavioral indicators of player experiences: Pressure patterns and postural responses. Presented at the Measuring

Behaviour 2008.

Inchauste, F. (2010, January 29). Better User Experience With Storytelling – Part One. *Smashing Magazine*. Retrieved from http://uxdesign.smashingmagazine.com/2010/01/29/better-user-experience-using-storytelling-part-one/

Infinity Ward. (2009). *Call of Duty: Modern Warfare 2*. Activision 2009.

Isbister, K., & Schaffer, N. (2008). *Game Usability: Advancing the Player Experience* (1st ed.). Morgan Kaufmann.

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, *66*(9), 641–661. doi:10.1016/j.ijhcs.2008.04.004

Kim, J. H., Gunn, D. V., Schuh, E., Phillips, B. C., Pagulayan, R. J., & Wixon, D. (2008). Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems. Presented at the CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. doi:10.1145/1357054.1357126

Kivikangas, M. J., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Jarvela, S., & Ravaja, N. (2011a). A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual Worlds*, *3*(3), 181–199. doi:10.1386/jgvw.3.3.181_1

Kivikangas, M. J., Nacke, L. E., & Ravaja, N. (2011b). Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. *ENTERTAINMENT COMPUTING*.

Korhonen, H., & Koivisto, E. M. I. (2006). Playability heuristics for mobile games. Presented at the MobileHCI '06: Proceedings of the 8th conference on Human-computer interaction with mobile devices and services, ACM Request Permissions. doi:10.1145/1152215.1152218

Lai-Chong Law, E. (2011). The measurability and predictability of user experience. Presented at the EICS '11: Proceedings of the 3rd ACM SIGCHI symposium on Engineering interactive computing systems, ACM Request Permissions. doi:10.1145/1996461.1996485

Lang, P. J. (1995). The emotion probe. *American psychologist*.

Laparra-Hernández, J., Belda-Lois, J. M., Medina, E., Campos, N., & Poveda, R. (2009). EMG and GSR signals for evaluating user's perception of different types of ceramic flooring. *International Journal of Industrial Ergonomics*, *39*(2), 326–332. doi:10.1016/j.ergon.2008.02.011

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). Research methods in human-computer interaction.

Lazzaro, N., & Keeker, K. (2004). What's my method?: a game show on games. *CHI EA "04: CHI "04 Extended Abstracts on Human Factors in Computing Systems*. doi:10.1145/985921.985922

Lewis, C., & Mack, R. (1982). Learning to use a text processing system (pp. 387–392). Presented at the the 1982 conference, New York, New York, USA: ACM Press. doi:10.1145/800049.801817

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Lin, T., Omata, M., Hu, W., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes? Presented at the OZCHI '05: Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future, Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

Livingston, I. J., Mandryk, R. L., & Stanley, K. G. (2010). Critic-proofing: how using critic reviews and game genres can refine heuristic evaluations. Presented at the Futureplay '10: Proceedings of the International Academic Conference on the Future of Game Design and Technology, ACM Request Permissions. doi:10.1145/1920778.1920786

Livingston, I. J., Nacke, L. E., & Mandryk, R. L. (2011). *Influencing Experience: The Effects of Reading Game Reviews on Player Experience*. *ICEC'11: Proceedings of the 10th international conference on Entertainment Computing* (Vol. 6972, pp. 89–100). Berlin, Heidelberg: ICEC 2011. doi:10.1007/978-3-642-24500-8_10

Lynn, J. (2013). Combining Back-End Telemetry Data with Established User Testing Protocols:

A Love Story (pp. 497–514). In Seif El-Nasr, M., Drachen, A., & Canossa, A. (Ed.), *Game Analytics: Maximizing the Value of Player Data*. London: Springer London. doi:10.1007/978-1-4471-4769-5_22

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, *5*(4), 333–369. doi:10.1016/S0364-0213(81)80017-1

Malone, T. W. (1984). Heuristics for designing enjoyable user interfaces: lessons from computer games. *Human factors in computer systems*.

Mandryk, R. L. (2008). Physiological Measures for Game Evaluation. In K. Isbister & N. Schaffer (Eds.), *Game Usability: Advancing the Player Experience*. Morgan Kaufmann.

Mandryk, R. L., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*.

Mandryk, R. L., Atkins, M. S., & Inkpen, K. M. (2006). A continuous and objective evaluation of emotional experience with interactive play environments. Presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. doi:10.1145/1124772.1124926

Marczak, R., van Vught, J., Nacke, L. E., & Schott, G. (2012). Feedback-based gameplay metrics: measuring player experience via automatic visual analysis. Presented at the IE '12: Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System, ACM Request Permissions. doi:10.1145/2336727.2336733

Marshall, C., & Rossman, G. B. (2010). *Designing Qualitative Research*. SAGE.

McAllister, G., & White, G. R. (2010). Video Game Development and User Experience. *Evaluating User Experience in Games*.

Medler, B., John, M., & Lane, J. (2011). Data cracker: developing a visual game analytic tool for analyzing online gameplay. Presented at the CHI '11: Proceedings of the 2011 annual conference on Human factors in computing systems, ACM. doi:10.1145/1978942.1979288

Medlock, M. C., Wixon, D., & Terrano, M. (2002). Using the RITE method to improve

products: A definition and a case study. *Usability Professionsals Association (2002)*.

Mirza-Babaei, P., & McAllister, G. (2011a). Biometric Storyboards: visualising meaningful gameplay events. Presented at the BBI Workshop CHI 2011, Vancouver.

Mirza-Babaei, P., & McAllister, G. (2011b). Biometric Storyboards to Improve Understanding of the Players' Gameplay Experience (pp. 1–10). Presented at the Inter-Disciplinary.net-Videogame Cultures and the Future of Interactive Entertainment, Oxford: Videogame Cultures and the Future of Interactive Entertainment 2011. Retrieved from http://www.inter-disciplinary.net/critical-issues/cyber/videogame-cultures-the-future-of-interactive-entertainment/project-archives/3rd/concurrent-session-6a-studying-gameplay/

Mirza-Babaei, P., Long, S., Foley, E., & McAllister, G. (2011). Understanding the Contribution of Biometrics to Games User Research. Presented at the Proceedings of DiGRA 2011, Proceedings of DiGRA.

Mirza-Babaei, P., Nacke, L. E., Fitzpatrick, G., White, G. R., McAllister, G., & Collins, N. (2012). Biometric storyboards: visualising game user research data. Presented at the CHI EA '12: Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts, ACM. doi:10.1145/2212776.2223795

Mirza-Babaei, P., Nacke, L. E., Gregory, J., Collins, N., & Fitzpatrick, G. (2013). How does it play better?: exploring user testing and biometric storyboards in games user research. Presented at the CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Request Permissions. doi:10.1145/2470654.2466200

Mirza-Babaei, P., Zammitto, V., Niesenhaus, J., Sangin, M., & Nacke, L. E. (2013). Games user research: practice, methods, and applications. Presented at the CHI EA "13: CHI "13 Extended Abstracts on Human Factors in Computing Systems, ACM. doi:10.1145/2468356.2479651

Moura, D., Seif El-Nasr, M., & Shaw, C. D. (2011). Visualizing and understanding players' behavior in video games: discovering patterns and supporting aggregation and comparison. Presented at the Sandbox '11: Proceedings of the 2011 ACM SIGGRAPH Symposium on Video Games, ACM Request Permissions. doi:10.1145/2018556.2018559

Nacke, L. E. (2009). *Affective Ludology :Scientific Measurement of User Experience in*

Interactive Entertainment.

Nacke, L. E. (2013). An Introduction to Physiological Player Metrics for Evaluating Games (pp. 585–619). In Seif El-Nasr, M., Drachen, A., & Canossa, A. (Ed.), *Game Analytics: Maximizing the Value of Player Data*. London: Springer London. doi:10.1007/978-1-4471-4769-5_26

Nacke, L. E. (2010, April 1). From Playability to a Hierarchical Game Usability Model. *arXiv.org*. doi:10.1145/1639601.1639609

Nacke, L. E., Bateman, C., & Mandryk, R. L. (2011). BrainHex: preliminary results from a neurobiological gamer typology survey. Presented at the ICEC'11: Proceedings of the 10th international conference on Entertainment Computing, Springer-Verlag.

Nacke, L. E., & Drachen, A. (2011). Towards a framework of player experience research. Presented at the EPEX ´11, Bordeaux.

Nacke, L. E., & Lindley, C. A. (2009). Affective Ludology, Flow and Immersion in a First-Person Shooter: Measurement of Player Experience. *Loading*, *3*(5).

Nacke, L. E., & Lindley, C. A. (2008). Flow and immersion in first-person shooters: measuring the player's gameplay experience. Presented at the Future Play '08: Proceedings of the 2008 Conference on Future Play: Research, Play, Share, ACM. doi:10.1145/1496984.1496998

Nacke, L. E., Drachen, A., Kuikkaniemi, K., Niesenhaus, J., Korhonen, H. J., Hoogen, W. M., et al. (2009). Playability and player experience research. *Proceedings of DiGRA*.

Nielsen, J. (1992). Evaluating the thinking-aloud technique for use by computer scientists. *(1992), pp. 69-82*, 69–82.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. Presented at the CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM. doi:10.1145/191666.191729

Pagulayan, R. J., & Steury, K. (2004). Beyond usability in games. *interactions*, *11*(5). doi:10.1145/1015530.1015566

Pagulayan, R. J., Keeker, K., Wixon, D., & Romero, R. L. (2003). User-centered design in games. Handbook for Human-Computer Interaction in Interactive Systems.

Pinelle, D., Wong, N., & Stach, T. (2008a). Heuristic evaluation for games: usability principles for video game design. Presented at the CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM. doi:10.1145/1357054.1357282

Pinelle, D., Wong, N., & Stach, T. (2008b). Using genres to customize usability evaluations of video games. Presented at the Future Play '08: Proceedings of the 2008 Conference on Future Play: Research, Play, Share, ACM. doi:10.1145/1496984.1497006

Poels, K., de Kort, Y., & Ijsselsteijn, W. A. (2007). "It is always a lot of fun!": exploring dimensions of digital game experience using focus group methodology. Presented at the Future Play '07: Proceedings of the 2007 conference on Future Play, ACM. doi:10.1145/1328202.1328218

Quesenbery, W., & Brooks, K. (2010). Storytelling for User Experience: Crafting Stories for Better Design, 1st edition. *Storytelling for User Experience: Crafting Stories for Better Design, 1st edition*.

Ravaja, N. (2004). Contributions of Psychophysiology to Media Research: Review and Recommendations. *Media Psychology*, *6*(2), 193–235. doi:10.1207/s1532785xmep0602_4

Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., & Keltikangas-Järvinen, L. (2008). The psychophysiology of James Bond: Phasic emotional responses to violent video game events. *Emotion*, *8*(1), 114–120. doi:10.1037/1528-3542.8.1.114

Relentless Software. (2009). *Buzz! Quiz World*. Sony Computer Entertainment Europe 2009.

Rosson, M. B., & Carroll, J. M. (2009). Scenario Based Design. In A. Sears & J. A. Jacko (Eds.), *Human-Computer Interaction: Development Process*. Human-Computer Interaction: Development Process.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502. doi:10.1037/0022-3514.57.3.493

Segal, K. R. (1991). Physiologic Responses to Playing a Video Game. *Archives of Pediatrics & Adolescent Medicine*, *145*(9), 1034–1036. doi:10.1001/archpedi.1991.02160090086030

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. *CHI EA "07: CHI "07 Extended Abstracts on Human Factors in Computing Systems*. doi:10.1145/1240866.1241057

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). Psychophysiological recording.

Sweetser, P., & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE*, *3*(3). doi:10.1145/1077246.1077253

The Economist. (2011a, December 10). All the world's a game. *The Economist*.

The Economist. (2011b, December 10). Thinking out of the box. *The Economist*.

The Economist. (2011c, December 10). Paying for pixels. *The Economist*.

van den Broek, E. L., Lisý, V., & Janssen, J. H. (2010). Affective man-machine interface: Unveiling human emotions through biosignals. *BIOSTEC 2009*, 21–47.

van Reekum, C., Johnstone, T., Banse, R., Etter, A., Wehrle, T., & Scherer, K. (2004). Psychophysiological responses to appraisal dimensions in a computer game. *Cognition & Emotion*, *18*(5), 663–688. doi:10.1080/02699930341000167

Vermeeren, A., Bouwmeester, den, K., Aasman, J., & de Ridder, H. (2002). DEVAN: A tool for detailed video analysis of user test data. *Behaviour & Information Technology*, *21*(6), 403–423. doi:10.1080/0144929021000051714

Voida, A., & Greenberg, S. (2009). Wii all play: the console game as a computational meeting place. *CHI*, 1559–1568. doi:10.1145/1518701.1518940

Voida, A., Carpendale, S., & Greenberg, S. (2010). The individual and the group in console gaming. *CSCW*, 371–380. doi:10.1145/1718918.1718983

Wallner, G., & Kriglstein, S. (2012). A spatiotemporal visualization approach for the analysis of

gameplay data. Presented at the CHI '12: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, ACM. doi:10.1145/2207676.2208558

Wallner, G., & Kriglstein, S. (2013a). PLATO: Understanding Gameplay Data Through Visualization. *GUR Workshop CHI 2013*, 1–5.

Wallner, G., & Kriglstein, S. (2013b). Visualization-based analysis of gameplay data: A review of literature. *ENTERTAINMENT COMPUTING*, *4*(3), 143–155. doi:10.1016/j.entcom.2013.02.002

Ward, R. D., & Marsden, P. H. (2003). Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies*, *59*(1-2).

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. doi:10.1037/0022-3514.54.6.1063

White, G. R., Mirza-Babaei, P., McAllister, G., & Good, J. (2011). Weak inter-rater reliability in heuristic evaluation of video games. Presented at the CHI EA '11: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, ACM. doi:10.1145/1979742.1979788

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental Psychology*. Oxford and IBH Publishing.

Yun, C., Shastri, D., Pavlidis, I., & Deng, Z. (2009). O“ game, can you feel my frustration?: improving user”s gaming experience via stresscam. Presented at the CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Request Permissions. doi:10.1145/1518701.1519036

Zammitto, V. (2011). The Science of Play Testing: EA's Methods for User Research. Presented at the Game Developers Conference 2011.

Appendix 1: Abbreviations and Acronyms

AI	Artificial intelligence
ANOVA	Analysis of variance: A statistics procedure for testing the fit of a linear model
ANS	Autonomic nervous system
BioSt	Biometric Storyboards
C++	An object-orientied general purpose programming language based on the C-language
CI	Confidence interval
CNS	Central nervous system
EDA	Electrodermal activity. Measurement of changes in ability of skin to conduct electricity
EEG	Electroencephalography. Measurement of brainwaves through electrodes on the scalp
EKG or ECG	Electrocardiography. Measurement of heart activity through skin electrodes
EMG	Electromyography. Measurement of muscle activity through electrodes on the skin
FPS	First-person shooters: A shooter game, where the player controls the viewport camera and weapons from first-person perspective
F2P	Free to Play
GEQ	Game Experience Questionnaire
GSR	Galvanic skin response: A certain type of EDA response
GUR	Games User Research
GURs	Games user researchers
HCI	Human-Computer Interaction
HD	High-definition
HR	Heart rate
HRV	Heart rate variability
IBIs	Interbeat intervals
M	Mean value
mV	Millivolts
MW2	Call of Duty: Modern Warfare 2
NA	Negative affect
NDA	Non-disclosure agreement
PA	Positive affect

PC	Personal computer
PNS or PeNS	Peripheral nervous system
PSNS	Parasympathetic nervous system
PX	Player experience
RITE	Rapid Iterative Testing and Evaluation
RM	Repeated measures
ROI	Return on investment
RPGs	Role-playing games
SAM	Self-assessment mannequin scale, see Lang (1980)
SC	Skin conductance, also referred to as EDA.
SCL	The level of skin conductance (SC) often described in the context of a phasic psychophysiological analysis
SD	Standard deviation
SDK	Software development kit. Compilation of tools, application and code libraries for the development of software
SNS	Sympathetic nervous system
SoNS	Somatic nervous system
TRUE	Tracking real-time user experience
UCD	User-centred design
UOIT	University of Ontario Institute of Technology
UT	User testing
UX	User experience
Wii	Nintendo Wii; a gaming console
µS	microSiemens

Appendix 2: Consent Form For S1, S2, S3 and Case Studies

CONSENT FORM FOR PROJECT PARTICIPANTS

PROJECT TITLE: **Video Game User Research (GUR)**

Project Approval Reference: **CREC-IEM/2012/01**

Dear participant,

You took part in the above Vertical Slice research project in collaboration with the University of Sussex. We have now contacted you to request a further permission in order to use the data we have collected from your gameplay session in the context of DPhil research. The data concerned are:

- A video of your gameplay session, sitting positions, physiological responses (such as Heart Rate and Skin conductance), verbal and interview comments about the game you played.

The footage will be reviewed by researchers at the University of Sussex and research collaborators. Anonymised data will be analysed and published as part of an Informatics thesis, and presented and published at research conferences and in journal articles.

Your identity will anonymised in any data shared with research collaborators, or presented in a publication. No identifying details will be made public in any way. Video footage showing your face or otherwise enabling identification will not be used in any public disclosure in any way.

This permission is voluntary, in that you can choose not to grant permission for part or all of the project, and you can turn down this request without being penalised or disadvantaged in any way.

In order to give permission, please tick the box below, sign and return this form:

☐ I consent to the processing of above data for the purposes of this research study. I understand that such information will be treated as strictly confidential and handled in accordance with the Data Protection Act 1998.

Name: _____

Signature _____

Date: _____

If you have any concern or questions please contact Pejman Mirza-Babaei: pm75@sussex.ac.uk

Appendix 3: Consent Form For S4

UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY
2000 SIMCOE STREET NORTH
OSHAWA, ONTARIO, CANADA L1H 7K4
T 905.721.8668 ext. 2830
F 905.721.3167
www.businessandit.uoit.ca
www.uoit.ca

UOIT
CHALLENGE INNOVATE CONNECT

FACULTY OF BUSINESS AND INFORMATION TECHNOLOGY

Informed Consent Form for Experiment.
This study has received ethical approval from the UOIT ethics committee (REB# 11-092)

5 July 2012

Investigators: Dr. Lennart Nacke, Faculty of Business and IT (Ext. 5356), Lennart.Nacke@uoit.ca
Pejman Mirza-Babaei, Faculty of Business and IT, Pejman.Mirza-Babaei@uoit.ca

This consent form is only part of the process of informed consent. Please print off this form for your personal records and reference. It should give you the basic idea of what this research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask your experimenter or any of the investigators listed above. Please take the time to read this form carefully and to understand any accompanying information.

This study is concerned with exploring the user experience and affective state on study participants during gameplay.

The goal of the research is to determine how the physiological (sensor-recorded) and self report (interview and survey-recorded) measures deviate from control levels when exposed to controlled media stimuli (primarily games and entertainment products).

About potential risks of silver-contact physiological sensors:
Physiological sensors use a silver contact area. These physiological sensors are harmless as they passively sense the electric conductivity of your skin. However, there low risks involved in this technique, such as a person getting a skin irritation from electrode gel or someone allergic to silver could have a skin reaction. Keep in mind that these are uncommon and not serious – to give you an example: In all the years that Dr. Nacke has worked with physiological sensors he has never encountered any of these reactions. If you experience any uncommon reaction during or after the experiment, please let the experimenter know. There are no other

UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY
2000 SIMCOE STREET NORTH OSHAWA, ONTARIO, CANADA L1H 7K4
T 905.721.8668 ext. 2830 F 905.721.3167
www.businessandit.uoit.ca www.uoit.ca

UOIT CHALLENGE INNOVATE CONNECT

FACULTY OF BUSINESS AND INFORMATION TECHNOLOGY

specific risks associated with the procedures and the equipment used in this study.

Your participation in this study is completely voluntary and you may interrupt or end the study at any time without giving a reason or fear of being penalized.

If at any point during the experiment you feel uncomfortable and want to end your participation, please let the experimenter know and they will end the study immediately.

The session will require about 30-60 minutes, during which you will be asked to play a game, read, and answer questions while your gameplay behavior will be logged using the computer and physiological sensors. You will watch your gameplay video and answer to interview questions.

At the end of the session, you will be given more information about the purpose and goals of the study, and there will be time for you to ask questions about the research.

As one way of thanking you for your time, we will be pleased to make available to you a summary of the results of this study once they have been compiled (usually within two months). This summary will outline the research and discuss our findings and recommendations. If you would like to receive a copy of this summary, please check the box below.

Thank you very much for your time and help in making this study possible. If you have any queries or wish to know more please contact Dr. Lennart Nacke, Faculty of Business and Information Technology, 2000 Simcoe St N, Oshawa, ON L1H 7K4. Phone: 905-721-8668 Ext. 5356 or email: Lennart.Nacke@uoit.ca

For any queries regarding this study, please contact the UOIT Research and Ethics Committee Compliance officer (compliance@uoit.ca and 905-721-8668 Ext. 3693.

UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY
2000 SIMCOE STREET NORTH
OSHAWA, ONTARIO, CANADA L1H 7K4
T 905.721.8668 ext. 2830
F 905.721.3167
www.businessandit.uoit.ca
www.uoit.ca

UOIT
CHALLENGE INNOVATE CONNECT

FACULTY OF BUSINESS AND INFORMATION TECHNOLOGY

After reading this information, you give consent.

- I understand that taking part in this study is my choice and that I am free to withdraw from the study at any time without reason and irrespective of whether or not payment is involved.
- This consent form will be kept in a locked filing cabinet in Oshawa for a period of seven years before being destroyed.
- I have read and understand all of the above information
- I understand that I am not waiving any of my legal rights

I, -

_____

(First Name, Last Name, signature), agree to take part in this research.

Voluntary and optional consent for photographic release

Please tick the following check box if you would like to give us photographic consent to use a video of you and the experimental setup in research reports and presentations.

☐ I would like to explicitly grant Dr. Nacke and his research assistants the right to use the video and audio material for presenting this study in publications, such as scientific journals and magazines, and research presentations. I understand that the video material is not linked to any personal data outside of this experiment that may identify me.

The non-visual data collected from this study will be used in research thesis, articles for publication in journals and conference proceedings. All data gathered is stored anonymously and kept confidential. Only Dr. Nacke and his research assistants may access and analyze the data. All published data will be coded, so that your non-visual data is not identifiable.

I, -

_____

(First Name, Last Name, signature), give consent to use video and image material of myself and the experimental setup in research reports and presentations.

Appendix 4: Interview Schedule for S3

Semi- structured Interview Plan

Stage 1: Introduction questions

Some ideas to ask about:

- Have you done user test on any of your games before? Or have you seen/given a user test report?
- Think about a recent case, what stage was the game in the development?
- What were you hoping to get from the report? Was it delivered?
- What was the form of the report?
- Who wrote the report and conducted user testing? Internal, publisher, external?
- How were findings presented?
- What did you like or find useful about the report? What do you think could be improve?
- How successfully do you think the report communicated the findings?
- What would you like to see in a typical user testing report?
- How would you improve user test report? How to better communicate its findings? What are the current limitations?
- What are the desires of game designers for a user testing report? What do you want to see in there? What do you need feedback on the most?

Stage 2: Show them BioSt prototypes and give them some time to explore them

Stage 3: Explore how they interpreted BioSt. Example of some possible questions to guide the discussion

- What is BioSt trying to communicate? (First impression)
- If they use actual storyboards in their game development? Would BioSt map to their game's storyboards?
- Ask them to point out some issues they think are more important to fix, and why?
- Can they use BioSt to get a sense of what needs to be done? And how to fix it?

Appendix 5: Player's Demographics Questionnaire

Your Age

Gender

Male

Female

Gaming Type

What type of gamer would you classify yourself as?

Non-gamer

Casual

Novice

Intermediate

Advanced

How often (on average) would you say that you play games?

Every day

A few times per week

A few times per month

Once per month

A few times per year

Once per year or less

Which mode of playing do you prefer?

Single player alone

Single player with other helping or controller/pad-passing

Multiplayer in the same room

Multiplayer over the Internet

Team/Cooperative play or clan play over the Internet

Virtual worlds or MMORPGs

Select your favourite gaming platforms?

You can select more than one

Nintendo, please specific:

Sony Playstation, please specific:

Microsoft XBOX, please specific:

Apple iOS platforms, please specific:

PC or Mac, please specific:

Mobile Phone, please specific:

Other:

Interests

Please select the genres that you enjoy playing

RTSs

FPSs

MMOs

RPGs

puzzlers

casual games

party games

racing games

story-led games

online games

sports games

music games

no preference

Other:

What is your all-time favourite game?

Notes

Please state anything else you may think is important that we know about you. Perhaps list your favourite games or how you heard about us.

Appendix 6: Questionnaire used in S4 - Game's Features

Please score the following features of the game:

		Very poor (1)	Poor (2)	Neutral (3)	Good (4)	Very good (5)
2.1	Jumps					
2.2	Time					
2.3	Score					
2.4	Controls					
2.5	Speed					
2.6	Collectables					
2.7	Level					

Appendix 7: Final Rating Questionnaire used in S4

Please rate the following overall aspects of the first game you have played:

		Very poor (1)	Poor (2)	Neutral (3)	Good (4)	Very good (5)
1	Experience					
2	Quality					
3	Fun					
4	Visuals					
5	Sounds					

Please rate the following overall aspects of the second game you have played:

		Very poor (1)	Poor (2)	Neutral (3)	Good (4)	Very good (5)
1	Experience					
2	Quality					
3	Fun					
4	Visuals					
5	Sounds					

Please rate the following overall aspects of the third game you have played:

		Very poor (1)	Poor (2)	Neutral (3)	Good (4)	Very good (5)
1	Experience					
2	Quality					
3	Fun					
4	Visuals					
5	Sounds					